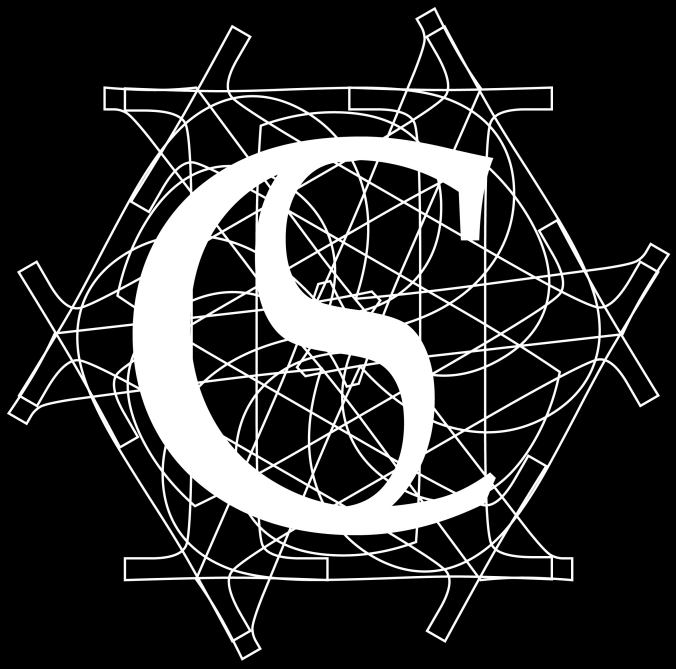


Essays in Politicial Economy

Carlo Rasmus Schwarz

July 2020

Thesis submitted for the Doctor of Philosophy in Economics

University of Warwick, Department of Economics

Contents

List of Figures

List of Tables

Acknowledgement

This doctoral thesis marks the end of my research at the University of Warwick. Locking back on the last 6 years, I am grateful for all the people who supported me and made my time in Warwick an enriching experience. First, my thanks go to my co-authors Mirko Draca, Thiemo Fetzer, Alessandro Iaria, Karsten Müller, and Fabian Waldinger. It was my great pleasure to work with you and learn from you. Without you, this dissertation and my research would not be of the same quality. Secondly, I am indebted to my friends and family with whom I spend countless hours of sublime happiness and who gave me the energy to continue my work. Lastly, I want to thank the many members of Warwick University and the economics academic community, who assisted me during my PhD.

Declaration

I hereby declare that the thesis is my own work based on collaborative research with Karsten Müller and Mirko Draca. In all cases, all authors contributed equally to the preparation and analysis of the data as well as the writing of the paper. The thesis does neither contain published material nor has it been submitted for a degree at another university.

Abstract

This dissertation consists of three papers investigating the causes and consequences of anti-liberal and populist trends that affected Western countries in the first two decades of the 20th century.

The first paper investigates the link between social media and hate crime in Germany. We show that anti-refugee sentiment on Facebook predicts crimes against refugees in otherwise similar municipalities with higher social media usage. To establish causality, we exploit exogenous variation in the timing of major Facebook and internet outages. Consistent with a role for "echo chambers", we find that social media posts contain narrower and more loaded content than news reports. Our results suggest that social media can act as a propagation mechanism for violent crimes by enabling the spread of extreme viewpoints.

The second paper studies whether social media can activate hatred of minorities, with a focus on Donald Trump's political rise. We show that the increase in anti-Muslim sentiment in the US since the start of Trump's presidential campaign has been concentrated in counties with high Twitter usage. To establish causality, we develop an identification strategy based on Twitter's early adopters at the South by Southwest festival. We also show that Trump's tweets about Islam-related topics are highly correlated with anti-Muslim hate crimes after the start of his presidential campaign, but not before.

The third paper sheds light on the trends in citizen polarization in Western countries. To this end, we propose a novel methodology to identify the underlying ideologies of citizens by applying Latent Dirichlet Allocation to political survey data. This approach indicates that in addition to a left-right scale, confidence in institutions defines another major ideological dimension. We find evidence for citizens shifting away from centrist ideologies into anti-establishment 'anarchist' ideologies over time. This trend is especially pronounced for the US.

Introduction

In 1992, Francis Fukuyama declared "The End of History" in his book of the same title (Fukuyama, 1992). He argued that the fall of the Soviet Union and thereby communism marked the final triumph of Western-style liberal democracy over other rival ideologies, such as monarchy and fascism. Liberal democracy supposedly marked the endpoint in the evolution of forms of government as it did not contain the inherent inconsistencies that led to the failure of previous ideologies. Fukuyama predicted that liberal consensus together with liberal values would spread around the globe reaching millions if not billions of people.

Over the last 30 years, these hopes have not yet turned out to be true. While liberal democracy could still triumph in the long run, in the current moment it rather appears that "History" has returned with a vengeance. Not only did liberal values not uniformly spread around the globe, authoritarian regimes are flourishing and consolidating their power in many countries (Economist Intelligence Unit, 2020). Most worryingly, even many Western countries have seen a backlash against the liberal world order, driven by a rise in populist parties and politicians (Norris and Inglehart, 2019), which oppose liberal values.

My PhD dissertation investigates the causes and consequences of some of these anti-liberal trends in three research projects. First, two co-authored papers with Karsten Müller study the effect of anti-minority sentiments on social media on real-life hate crimes. These projects are motivated by the fact that many populists appear extremely successful in utilizing social media for their political purposes (Sunstein, 2017). While speeches of politicians always have been understood as an important policy tool, the advent of social media increased the speed and breadth with which such information is disseminated. Recent work by Guriev et al. (2019) provides further evidence for the fact that populists appear to profit from the rollout of 3G mobile broadband across countries.

Among all the consequences of these new media platforms on society (e.g. polarization and misinformation), their link to violence is the most worrisome, as it not only undermines the government's monopoly on the use of force (Weber, 1919) but also threatens cooperation in civil society. While the potential consequences of hateful online rhetoric and offline outcomes have been widely debated in the media and among policymakers, there is little empirical evidence for a causal relationship between the two phenomena. Karsten Müller and I attempt to close this gap with the two research projects.

1) Fanning the Flames of Hate: Social Media and Hate Crime

This paper studies the link between social media and hate crime using data from the Facebook page of the right-wing party "Alternative für Deutschland" (AfD). The AfD as the first

far-right party in the German parliament since 1945, is the most important representative of the new populist trends in Germany. Particularly during the European refugee crisis in the years 2015 and 2016, the AfD's Facebook page became a hub for the exchange of anti-refugee messaging. The paper uses this fact to create a measure of anti-refugee sentiment in German social media, which as we show, predicts violence against refugees, especially in municipalities with higher exposure to such content. To establish causality, the paper puts forward a novel identification strategy based on local internet outages and Germany-wide Facebook outages to create quasi-experimental variation in Facebook exposure. We document that even these short-run interruptions to social media are associated with decreases in the number of anti-refugee incidents. Together with additional evidence on the high prevalence and negative tone of anti-refugee speech on far-right social media, the findings of this paper suggest that social media are indeed a powerful tool for the transmission of anti-minority sentiments. As such, our work contributes to the existing literature on media and violence, which so far focused on nationalistic propaganda in settings of high ethnic tensions Yanagizawa-Drott (2014); Adena et al. (2015) and DellaVigna et al. (2014). In contrast, our setting highlights that even in absence of state-sanctioned anti-minority propaganda, social media provides an alternative forum for the exchange and spread extreme rhetoric and viewpoints for the fringe elements of society.

2) From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment

The second paper with Karsten Müller sheds additional light on the mid- to long-term consequences of the exposure to social media content. In particular, we focus on the 45th president of the United States, Donald J. Trump, one of the most prominent representatives of recent anti-liberal trends. We ask whether Donald Trump's political rise was associated with a rise in hate crimes against Muslims and if anti-Muslim sentiments propagated through the social media platform Twitter.

In line with this hypothesis, we first show that the increase in anti-Muslim hate crime since the start of Trump's presidential campaign has been concentrated in counties with high Twitter usage. To rule out that other county characteristics that correlate with Twitter usage could explain this finding, we develop an identification strategy based on Twitter's early adopters at the South by Southwest (SXSW) festival. The 2007 SXSW festival marked a turning point in Twitter's popularity. We show that Twitter usage picked up considerably after the festival in areas with more SXSW 2007 attendees and that these areas still have higher Twitter usage today. People who started following SXSW on Twitter prior to the 2007 event have no such predictive power, nor do other festivals similar to the SXSW festival.

Using the likely 2007 SXSW attendees as an instrument for Twitter usage, we confirm that social media played a role in the increases in anti-Muslim hate crime.

We also show that Trump's tweets about Islam-related topics are highly correlated with spikes in anti-Muslim hate crimes after the start of his presidential campaign, but not before. To rule out that such spikes are driven by other events that occur at the same time (e.g. terrorist attacks), we exploit that Trump is more likely to tweet about Muslims on days when he plays golf in an instrumental variable framework. This analysis is motivated by the fact that many commentators have argued that golf shifts Trump's state of mind away from politics. We find that our findings also persist when instrumenting for Trump's anti-Muslim tweets. As further evidence of the mechanism, we illustrate how Trump's tweets lead his Twitter followers to share his messages and produce their own anti-Muslim content. Taken together, the evidence suggests that Trump's presidential campaign and Twitter rhetoric influenced the willingness of some people to commit hate crimes against Muslims, in the short-run as well as in the mid- to long term. The paper also provides important insides into the importance of populist leaders and their support for existing anti-minorities views.

3) How Polarized are Citizens? Measuring Ideology from the Ground-Up

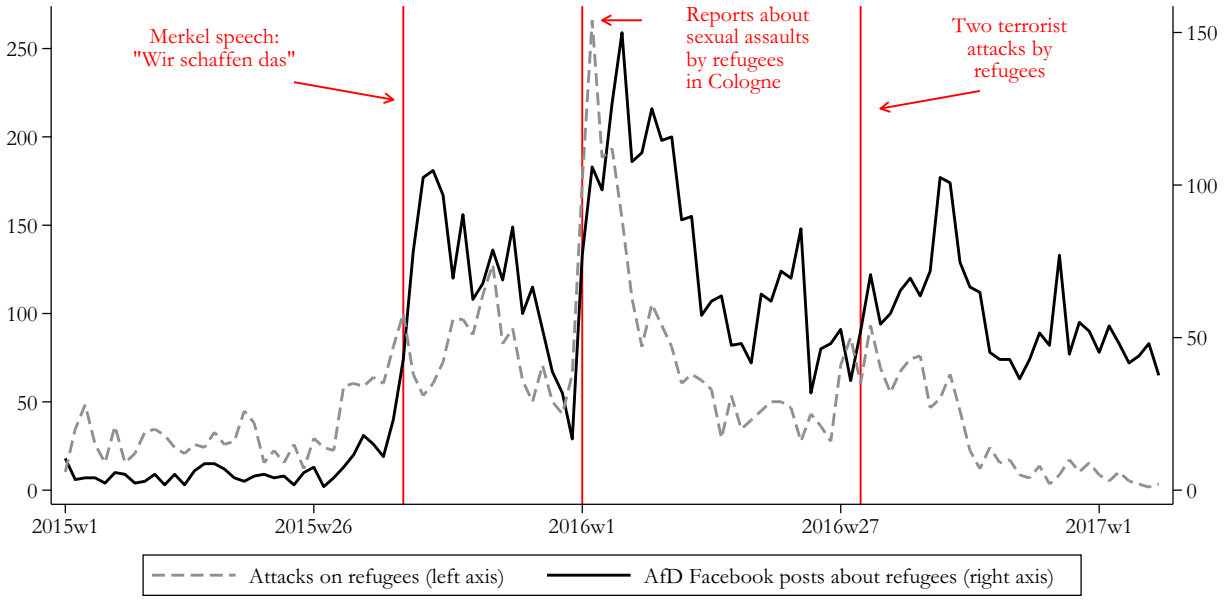
The last paper in my dissertation makes a more conceptual and methodological contribution to our understanding of populist movements and polarization. My joint work with Mirko Draca proposes a new methodology to measure the underlying ideologies of citizens based on the unsupervised machine learning technique Latent Dirichlet Allocation. We exploit the fact that people of similar ideologies give similar answers to the same questions. Latent Dirichlet Allocation can then identify the underlying ideologies that best describe the responses in the data. An ideology is then simply defined as particular questions response profile. Each individual, in turn, is described as a probabilistic mixture of ideological types.

This approach hence allows us to identify the ideologies of citizens exclusively based on the co-occurrence patterns of responses in the World Value Survey. This approach has 3 main advantages. First, it allows us to move beyond pre-defined definitions of ideology like the left-right spectrum or party affiliation. Secondly, it allows for a comparison off ideologies across countries and over time. Third, given the flexibility of our approach, we believe it to be applicable to a wide array of other settings in which researchers so far relied on predefined groups.

Our analysis uncovers, in addition to the traditional left-right spectrum, the existence of a stable ideology centered around the distrust in major societal institutions (e.g. parliament, churches). It is this ideology that drives support for populist parties. As such our work provides a rare cross-country perspective in the growth of populist movements in Western countries.

Further, the created individual level type shares allow us to investigate the polarization of citizens based on the measures suggested by Esteban and Ray (1994); Duclos et al. (2004). We again find that the increasing prevalence of low trust ideologies drives recent increases in the polarization of citizens, especially in the United States.

Together the three papers in my dissertation expand our understanding of modern populist movements. The papers simultaneous highlight roots and dangers of populism which appear to emerge independently from the country under investigation.



1) Fanning the Flames of Hate: Social Media and Hate Crime

Carlo Schwarz (University of Warwick)

Karsten Müller (Princeton University)

1.1 Introduction

Social media has come under increasing scrutiny in recent years. In the wake of the 2016 presidential election in the United States, for example, relatively recent phenomena such as fake news, social media echo chambers, and bot farms have been subjects of widespread media coverage and public discourse (e.g Times, 2016, 2017b). The role of online hate speech in particular has been at the center of an intense and polarized debate. Despite public interest and calls for policy action, there is little empirical evidence on how hateful social media content translates into real-life behavior.

In this paper, we investigate the role of social media in the propagation of hate crimes. Previous research has shown that traditional media can play a role in violent outbursts or ethnic hatred (e.g Yanagizawa-Drott, 2014; Adena et al., 2015; DellaVigna et al., 2014). In contrast to traditional media, social media platforms allow users to easily self-select into niche topics and extreme viewpoints. This preferential selection may limit the spectrum of information people absorb and create "echo chambers" (Sunstein, 2009, 2017), which reinforce similar ideas (see e.g. Bessi et al., 2015; Del Vicario et al., 2016; Schmidt et al., 2017). Social media has also become a widely-consumed news source, particularly for young people: in Germany, for example, social media is among the main news sources of 18 to 25 year olds (Hölig and Hasebrink, 2016). In the US, around half of all adults use social media to get news and two thirds of Facebook users use it as a news source (Center, 2018). This suggests that social media could be particularly effective in propagating hateful sentiments.

We study the link between anti-refugee sentiment on Facebook and hate crimes against refugees in Germany. The German setting is motivated by the influx of around one million refugees into the country between 2015 and 2016 (BAMF, 2016), which was accompanied by frequent violent crimes committed against them (see, for example, recent video coverage by Times, 2017c). Between January 2015 and early 2017 alone, the non-profit organization "Amadeu Antonio Stiftung" recorded around 3,300 anti-refugee incidents, including over 750 cases of arson or outright assault.

We posit that social media can reinforce anti-refugee sentiments, which may push some potential perpetrators over the edge to carry out violent acts. Our empirical strategy exploits differences in Facebook usage at the municipal level and weekly variation in anti-refugee sentiment on social media. We create a novel measure for the salience of anti-refugee hate speech on social media based on the Facebook page of the "Alternative für Deutschland"

(Alternative for Germany, AfD hereafter), a relatively new right-wing party that became the third-strongest faction in the German parliament following the 2017 federal election. The AfD has positioned itself as an anti-refugee and anti-immigration party. With more than 300,000 followers, 175,000 posts, 290,000 comments, and 500,000 likes (as of early 2017), their Facebook page has a broader reach than that of any other German party.[1]

This widespread reach makes the AfD's Facebook page uniquely suited to measure anti-refugee sentiment on social media. In contrast to established political parties like Angela Merkel's Christian Democratic Union (CDU) or the German Social Democrats (SPD), the AfD allows users to directly post messages on its Facebook wall. The AfD is also the only party that does not explicitly outline rules of conduct, e.g. by threatening to remove racist, discriminating, or otherwise hateful comments. We show that the content on the AfD page is consistently more focused on refugees than that of traditional news reports and frequently contains loaded terms that civil rights groups have identified as "hate speech". These detailed data also allow us to construct a measure of each town's exposure to Germany-wide anti-refugee sentiment using the share of the population that is active on the AfD Facebook page.

Using fixed effects panel regressions, we find that—during periods of high salience of refugees on right-wing social media—anti-refugee hate crimes increase in areas with higher Facebook usage. This correlation is especially pronounced for violent incidents such as assault. Controlling for a large vector of municipality characteristics, interacted with our salience measure, makes little difference for the magnitude and statistical significance of these estimates.

A concern is that our measures of exposure to right-wing social media may be correlated with unobserved municipal characteristics that explain disproportionate increases in hate crimes during times of high anti-refugee sentiment. To narrow down the social media transmission channel, we provide quasi-experimental evidence using the exact timing of country-wide Facebook outages and local internet disruptions, which reduce the number of social media posts.

To begin, we study large, Germany-wide Facebook outages resulting from programming or server problems at the platform. These outages disrupt users' exposure to this particular social media platform without affecting other online channels. We find that Facebook disruptions reduce local hate crimes, particularly in areas with many AfD users. Further,

[1]We provide a short history of the AfD in Section 1.5.1 in the online appendix.

during Facebook outages, higher anti-refugee sentiment is not associated with a differential increase in hate crimes in areas with high Facebook usage. These results suggest that social media might play a propagating role in translating online content into offline violence.

We also exploit the precise timing of hundreds of local internet disruptions as a source of granular exogenous variation in access to social media. These local disruptions reduce a particular town's exposure to social media content while leaving Germany-wide refugee salience unaffected. Notably, the frequency of internet disruptions is geographically dispersed and largely unrelated to observable local characteristics, including AfD likes on Facebook.

We find that, while hate crimes increase in periods of higher refugee salience, this correlation disappears for municipalities experiencing an internet outage. Quantitatively, a typical internet disruption fully mediates the link between social media and hate crime. Further, once we take into account social media transmission, these internet outages themselves are no longer associated with anti-refugee incidents, nor are their interactions with local internet usage or mobile internet access. These results point to social media as propagation mechanism rather than other online channels. It also makes it unlikely that we are capturing a "displacement effect" arising from potential perpetrators fixing their internet access.

We also analyze how other salient news events mediate the link of anti-refugee Facebook posts with the number of violent incidents, building on Eisensee and Strömberg (2007) and Durante and Zhuravskaya (2018). Specifically, we look at the European Soccer Championship, Brexit, and Donald Trump's presidential election, all of which crowded out the salience of refugees. Similar to our outage results, social media exposure has a significantly more muted relationship with hate crimes during these events. The link we uncover appears to be specific to anti-refugee sentiment: other posts on the AfD Facebook page, e.g. those related to Muslims or the European Union, do not have the same predictive power for anti-refugee hate crimes. Consistent with the hypothesis that social networks can act as transmission channel, the correlation with hate crime is larger in regions where AfD users show higher Facebook engagement.

When interpreting our results, we do not claim that social media itself causes crimes against refugees out of thin air. Rather, our argument is that social media can act as a propagating mechanism for hateful sentiments that likely have many fundamental sources. We find evidence for two potential channels. First, our results are driven by refugee attacks committed by groups of perpetrators. This suggests that social media may motivate collective action, consistent with existing evidence on other political outcomes such as protests (e.g.

Enikolopov et al., 2020). Second, we find evidence for a spillover channel. Hate crimes are considerably more common in weeks when neighboring towns also experience them, and this is particularly true for towns with many right-wing social media users when anti-refugee sentiment is elevated. In contrast, we find little evidence that social media provides useful information to perpetrators. Our results are also unlikely to be explained by persuasion effects, because we focus on high-frequency variation.

Related literature. Our work provides evidence that social media may have effects on real-life outcomes, as measured by hate crimes. We build on existing work on media exposure and persuasion (see e.g. DellaVigna and Gentzkow, 2010; DellaVigna and La Ferrara). In addition to work on traditional media and violence cited above, Dahl and DellaVigna (2009) show that—in contrast to experimental settings—violent movies decrease violent crime in the field due to displacement effects. Television has also been associated with short-lived outbursts of domestic violence (Card and Dahl, 2011). In other research, Bhuller et al. (2013) demonstrate that exposure to pornographic material on the internet is linked to increased sex crime. Bursztyn et al. (a) find that media coverage of close elections increases voter turnout, while Gavazza et al. (2018) show that broadband diffusion decreased voter turnout in the United Kingdom (see also Gentzkow, 2006; Manacorda and Tesei). Enikolopov et al. (2020) find that social media exposure spurs protest participation in Russia by reducing coordination costs.

We contribute to this literature by investigating the role of social media in stirring up violence. Previous research has documented the prevalence of online hate speech (Oksanen et al., 2014). Other work has shown that Google search data can be used to measure racial animus (Stephens-Davidowitz, 2014). In complementary work, we study the effect of Twitter usage on anti-minority sentiments in the United States (Müller and Schwarz, 2018b). Bursztyn et al. (b) study the effect of social media on xenophobia in Russia. In contrast to these papers, we focus on the short-run impact of social media posts, rather than long-run effects that may work through persuasion or changes in social norms.

Our paper also builds on research about the polarization of citizens (e.g Fiorina and Abrams, 2008). There is no consensus on whether social media increases or decreases polarization: some authors argue that social media are divisive (Pariser, 2011; Gabler, 2016), while others find that polarization *decreases* with social media usage (Barberá, 2014; Boxell et al., 2017). Our work suggests that—independent of whether social media affects overall polarization or not—social media content can be associated with violent crimes.

We also build on the literature on culture and violence. Summarizing a vast body of research, Alesina and La Ferrara (2005) find that cultural and and religious fragmentation predict the likelihood of civil war across countries. Voigtlander and Voth (2012) show that anti-Semitic violence in Germany is highly persistent: pogroms during the era of the Black Death predict pogroms in the 1920s, Jewish deportations, and synagogue attacks during the rise of the Nazi party. Similarly, Jha (2013) shows that medieval interethnic complementarities in trade decrease the likelihood of modern Hindu-Muslim riots. These papers, however, are largely silent on the existence of volatile, short-lived bursts of sentiment leading to violent incidents. As such, our work is also related to Fouka and Voth, who show that monthly variation in public acrimony between Greek and German politicians during the Greek debt crisis affected German car purchases particularly in areas of Greece where German troops committed war crimes during World War II. Our results also align with the findings of Colussi et al. (2016), who show that a higher salience of minority groups increases the likelihood of hate crimes.

While traditional media such as television are regulated in most countries, legislators are now beginning to address social media. Our work is thus particularly topical in light of the political discussions in many countries about anti-hate speech laws and censoring hate speech on social media. The German parliament, for example, passed an anti online hate speech law ("Netzwerkdurchsetzungsgesetz") on June 30, 2017, which threatens providers of online platforms such as Facebook with fines up to EUR 50 million for failing to delete "criminal" content that is "obviously unlawful". The controversial law was the initiative of German Minister of Justice Heiko Maas, who lamented social media platforms' unwillingness to address "online hate crime".[2] The European Union has issued independent guidelines calling on social media companies to remove illegal hate speech as well. In the United Kingdom, the Crown Prosecution Service plans to increase prosecution of online hate crimes (Guardian, 2017; BBC, 2017). Our paper serves as a first attempt to address this important topic empirically.

The paper proceeds as follows. In Section 1.2 we introduce the data used in our empirical analysis. Section 1.3 presents the results. Section 1.4 concludes.

[2]See, for example, the official statement of the German parliament on bundestag.de.

1.2 Data

We construct a dataset on social media activity and anti-refugee hate crimes in Germany. In total, we combine data from 12 different sources which we describe in more detail in the following subsections: (1) Municipal-level data on anti-refugee hate crimes; (2) Facebook data on posts, likes, and comments on the AfD page; (3) hand-collected municipal-level data on Facebook user locations; (4) municipal-level data on internet outages; (5) a hand-coded dataset on major weekly Facebook outages; (6) municipal- and county-level socioeconomic data from the German Statistical Office; (7) municipal-level voting data; (8) county-level data on broadband access; (9) municipal-level data on newspaper sales; (10) data on the content of reporting about refugees from Nexis; (11) city-level data on neo-Nazi murders and historical anti-Semitism; and (12) weekly Google search data on major news events in our sample. The final panel dataset covers 4,466 German municipalities for the 111 weeks from 1st January 2015 to 13th February 2017. Summary statistics for the main variables of interest can be found in Table 1.1 and Table 1.10. The online appendix provides a comprehensive overview of the data sources and variable definitions (see Table 1.11).

1.2.1 Anti-Refugee Incidents

The data on incidents targeting refugees were collected by the Amadeu Antonio Foundation and Pro Asyl (a pro asylum NGO).[3] These data cover incidents including anti-refugee graffiti, arson of refugee homes, assault, and incidents during protests in Germany between January 2015 and early 2017. This period is of particular interest since it includes the beginning and height of the refugee crisis in Germany. All 3,335 anti-refugee aggressions feature a short description and are classified into four groups. The most common cases are property damage to refugee homes (2,226 incidents), followed by assault (534), incidents during anti-refugee protests (339), arson (225). 11 events are classified as suspected cases that were still under investigation. Table 1.9 in the online appendix lists examples for each class of anti-refugee activity.

All incidents are geo-coded with an exact longitude and latitude, which we use to assign them to municipalities.[4] Figure 1.1 shows the location of the anti-refugee incidents in our

[3]These data are available at `https://www.mut-gegen-rechte-gewalt.de/service/chronik-vorfaelle`.

[4]To assign coordinates to municipalities, we use the shape files provided by the ©GeoBasis-DE/BKG 2016 website. The shape file contains data for the 4,679 German municipalities ("Gemeindeverwaltungsverband").

Table 1.1: Summary Statistics for Main Variables

	Level	Obs	Mean	SD	Min.	Max.
Refugee Attacks						
Refugee attacks	Muni.-Week	495,726	0.007	0.099	0	8
Arson attacks	Muni.-Week	495,726	0.000	0.022	0	2
Other property damage	Muni.-Week	495,726	0.004	0.076	0	8
Assaults	Muni.-Week	495,726	0.001	0.035	0	3
Protests	Muni.-Week	495,726	0.001	0.030	0	5
Social Media Data						
AfD users/Pop.†	Municipality	495,726	0.301	0.286	0	8
Refugee posts	Week	495,726	84	61	2	259
Posts/AfD users	Municipality	395,493	0.554	3.882	0	118
Comments/AfD users	Municipality	395,493	1.1	7.3	0	270
Likes/AfD users	Municipality	395,493	1.8	12.3	0	370
Auxiliary Variables						
$I_{Internet\,outage}$	Muni.-Week	495,726	0.001	0.025	0.000	1.000
$I_{Facebook\,outage}$	Municipality	495,726	0.072	0.259	0.000	1.000
Baseline Controls						
Ln(Population (2015))	Municipality	495,726	9	1	6	15
GDP/Worker	County	493,617	63,095	9,846	46,835	136,763
Population density	Municipality	495,726	282	382	7	4,653
AfD vote share (2017) (in %)	Municipality	492,618	15	7	3	45
Share high school (in %)	Municipality	495,726	29	8	0	58
Share broadband access (in %)	Municipality	495,726	83	11	44	100
Share immigrants (in %)	Municipality	483,072	14	8	2	50
Asylum Seekers/Pop.	County	495,726	0.011	0.006	0.000	0.102

Notes: This table reports summary statistics for the main variables in the estimation sample. Variables tagged with a † are scaled by population (in 1,000).

observation period for each German municipality.

The data appear to be high quality. Each entry has a clearly indicated source. Nearly half of the incidents in the dataset are reported by the federal government in response to inquiries by the left-wing party "Die Linke". Other sources include police reports and national or local media outlets. We hand-checked a random sample of 100 incidents and found their coding accurately reflected the information reported in the respective source.

1.2.2 Facebook Data on Refugee Salience

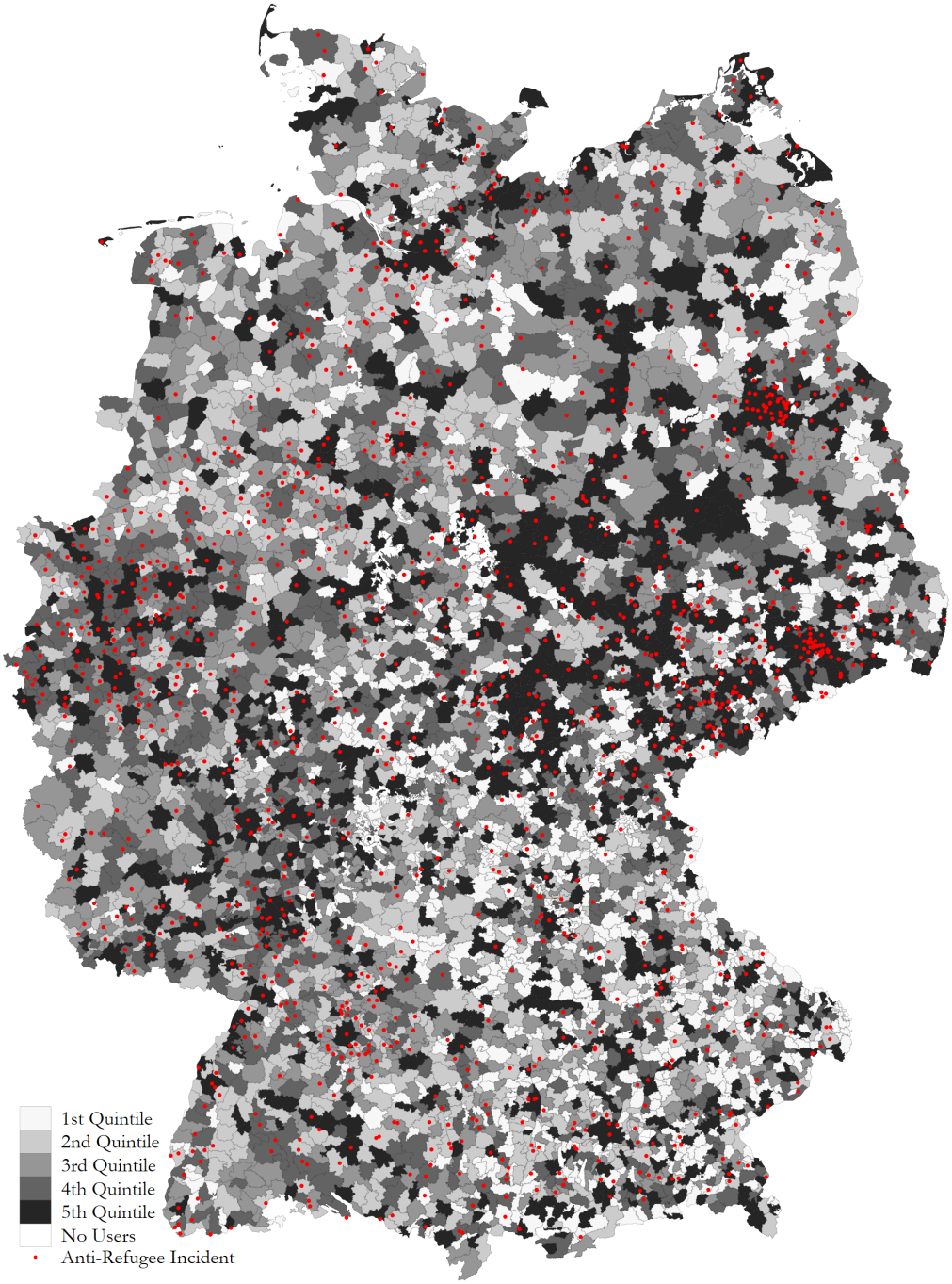
We construct a proxy for the frequency of anti-refugee messaging on social media based on the Facebook page of the AfD. We chose the AfD's page because the party is by far the most popular far-right political movement in Germany. At the time of the refugee crisis, the AfD also had the highest number of Facebook followers of *any* German party. This makes their page arguably the most important platform of exchange about refugees among Germany's right-wing social media users.

We start by using the Facebook Graph API to collect all status posts, comments, and likes from the AfD Facebook page (see Section 1.5.2 for an introduction to Facebook). The API provides a unique identifier for each post, allowing us to link posts to comments and likes, as well as the users who posted, commented, or liked anything on the page. Overall, we collected 176,153 posts, 290,854 comments, 510,268 likes, and 93,806 individual user IDs.

As our baseline measure for the salience of anti-refugee hate speech on social media, we use the number of posts on the AfD Facebook page that contain the word "Flüchtling" (refugee) in any given week. The narrative in these posts centers around the idea that the "elites"—politicians and mainstream media outlets—have betrayed "the people" by allowing "streams" of illegitimate "economic refugees" to enter the country, who are described as being criminals and rapists for "cultural reasons". Table 1.8 in the online appendix provides a few representative examples; Section 1.3.5 provides a more in-depth analysis. A potential downside of this approach is that we may inadvertently tag posts that do not express negative sentiments towards refugees. However, a careful content analysis of posts and comments reveals that the overwhelming majority appear to agree with the positions of the AfD. This

213 of these municipalities do not have inhabitants (e.g. forest areas) nor anti-refugee incidents. After dropping these cases, we are left with 4,466 municipalities in our estimation sample. We use the level of the "Gemeindeverwaltungsverband" since these exhibit smaller differences in their size and population than the 11,165 German "Gemeinden" and are therefore more suitable for spatial analysis according to the data provider (see link).

Figure 1.1: AfD Facebook Usage per Capita and Anti-Refugee Incidents

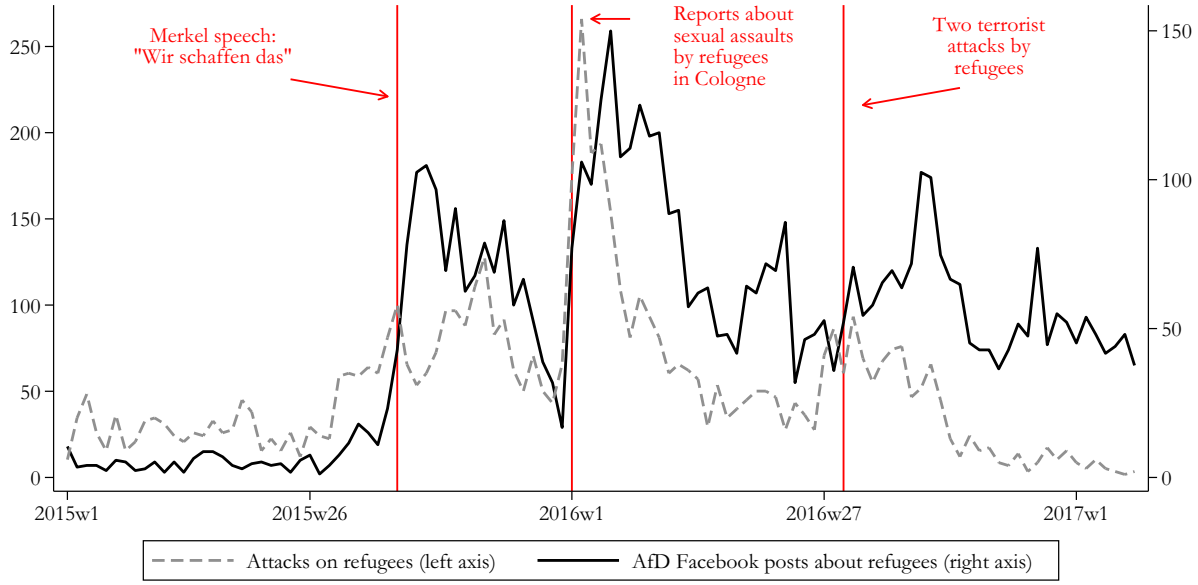


Notes: This map plots the number of Facebook users of the Alternative for Germany (AfD) page per capita for each of the 4,466 German municipalities. The red dots indicate the locations of the 3,335 anti-refugee incidents from the Amadeu Antonio Foundation.

is perhaps unsurprising given that only people who "like" the AfD Facebook page will be informed about new posts. Critics, on the other hand, have a strong incentive not to indicate publicly that they "like" the party.

We plot the total number of AfD Facebook page posts about refugees and the number of anti-refugee incidents in Figure 1.2. Weeks with more refugee posts also tend to have more anti-refugee events. Both series clearly spike during salient events related to refugees, such as Angela Merkel's widely reported statement "Wir schaffen das" ("We can do this") during a press conference on the challenges of the refugee situation. A simple time series regression of refugee attacks on AfD posts yields a R^2 of 0.34 (unreported).

Figure 1.2: Refugee Posts on Social Media and Anti-Refugee Incidents Over Time



Notes: This figure plots the number of posts about refugees on the Facebook page of the "Alternative for Germany" and the number of anti-refugee incidents in Germany over time.

1.2.3 Municipal-Level Facebook Measures

We construct a measure of exposure to right-wing social media at the municipal level. Because survey data about German Facebook usage are, to our knowledge, only available at the level of the 16 federal states, we hand-collect user location data by using the unique user identifiers provided by the Facebook Graph API. Due to Facebook's privacy policy, we are only able to collect this information for people who make it publicly available.

Because we are interested in the transmission of right-wing social media sentiment, we measure exposure to it on Facebook based on users of the AfD page. In total, we can identify 93,806 users who interacted with the page at least once.[5] We were able to hand-collect and geocode a place of residence for 34,396 of these users. Overall, we were able to identify at least one AfD Facebook page user for 3,563 of the 4,466 municipalities.[6] In Figure 1.1 we visualize the distribution of AfD users per capita. Anti-refugee incidents are concentrated in areas with more right-wing social media users. To illustrate this, Figure 1.9 in the online appendix shows the share of municipalities with at least one refugee attack, depending on whether we can identify *at least one* AfD Facebook page user. Municipalities with AfD users are three times as likely to experience an attack during our observation period. Out of the total 3,335 attacks on refugees in our sample, 3,171 occurred in municipalities with AfD Facebook page users. A t-test rejects the null hypothesis of no difference between the mean of the two groups with a value of 22.95.

Using the location data for AfD users, we can also assign posts, comments, and likes to municipalities. Based on these data, we construct auxiliary measures of social media interactions, e.g. the number of local posts scaled over the number of AfD users.[7]

1.2.4 Data on Internet and Facebook Outages

We collect data on local internet outages from Heise Online. Heise lists user reports of internet problems by telephone area codes and includes start times and duration. We use area codes to assign internet problems to municipalities; the start date and duration allow us to count the number of problems for each municipality and week.[8] The internet outage reports are geographically dispersed with no clear patterns of regional clustering (see Figure 1.11a). The outages are also dispersed over time Figure 1.11b.

[5]The Facebook API does not provide data on which users "like" a page but only on users who *interact* with a page, e.g. by liking another user's comment. As a result, the total number of user IDs we have is smaller than the more than 300,000 people who had liked the AfD Facebook page as of 2017.

[6]Note that the decision of users to disclose their location is unlikely to matter in our setting. This is because we exploit variation *within* the same location over time, which abstracts from time-invariant endogenous selection using municipality fixed effects.

[7]We find that some users post and comment excessively, which leads to a few outliers in measuring how active users are in a given municipality. We therefore winsorize the number of posts, comments, and likes we can attribute to local users at the 99.9th percentile to avoid individual users driving the results.

[8]If an area code spans multiple municipalities, we assign an internet outage to the municipality that overlaps most with the area code. We prefer this over to assigning the outage to all municipalities within the area code's territory because some area codes include minor overlaps with many municipalities. Assigning an internet outage to all of these municipalities would introduce substantial noise.

To validate the Heise data, we search for newspaper reports on major internet disruptions. While the large-scale and short-lived outages discussed in the newspaper reports are not representative of the more localized and longer-lasting outages we exploit in our regressions, they do suggest that the Heise data provide a valid proxy for internet disruptions. For all major disruptions we could identify in newspapers, the Heise data suggest an increase in the number of outages specific to the internet provider experiencing the outage. Table 1.12 lists several examples of newspaper reports on such outages and the respective information in our data.[9]

We focus on major outages that fulfill two criteria: (1) they have to last longer than 24 hours, and (2) they affect a significant part of the population (be in the top quartile of the reported internet problems to population ratio). This gets around the issue that some reports may reflect individual users' glitches rather than general disruptions.[10]

We also collect information on major Facebook disruptions. To identify these, we start by searching for newspaper reports of Facebook problems in our sample period. In total, we find reports on eight large outages (see Table 1.13 for an overview and more details). We then validate their precise timing using the number of weekly user-reported Facebook problems on the website of "Allestörungen", a portal for aggregating user complaints on individual websites and apps. Perhaps unsurprisingly, the eight outages widely reported on in the news media are also associated with spikes in user-reported problems.

Using these data, we define a dummy variable that is 1 for weeks with Facebook outages and 0 otherwise. These outages have the advantage that they are specific to Facebook; in fact, they are uncorrelated with the total number of weekly internet outages in a given week from our Heise data. In contrast to the internet disruptions, the downside is that Facebook outages are rare, shorter, and only generate weekly variation.

1.2.5 Auxiliary and Control Variables

We obtain control variables from a host of sources, which are explained in more detail in the online appendix. Socioeconomic data on the municipality and county level are from the

[9] To interpret the number of outages, note that the Heise data reports an average of four reported internet outages per provider per week. That means even an increase of 15 reported outages represents a large increase.

[10] In some cases, users do not seem to report the end date of the internet outage, which can lead to unlikely durations of several months. We thus winsorize the maximum duration at 3 weeks, but this choice is not material for our results. We scale outages over population because towns with more inhabitants mechanically also report more disruptions. As we discuss below, our results are robust to using alternative definitions of this cut-off.

German Statistical Office, available via www.regionalstatistik.de. We include information on each municipality's population by age group, GDP per worker, population density, the share of the population with a high school degree ("Abitur"), the share of the population receiving social benefits, the share working in manufacturing, and the vote results for the 2017 German Federal Election. To control for "pull factors" of anti-minority crimes, we also obtain the share of the population that are immigrants and asylum seekers.

To measure the extent to which people use the internet, we use the share of households in a county with broadband access as well as average mobile download speeds, collected by the Federal Ministry of Transport and Digital Infrastructure (BMVI).[11] In addition, we use the number of registered *.de* internet domains per capita in a county to measure internet affinity, which has a correlation of 0.48 with broadband access.

To measure the local penetration of traditional media, we obtain data for 2016/2017 newspaper sales from the "Zeitungsmarktforschung Gesellschaft der deutschen Zeitungen (ZMG)" (Society for Market Research of German Newspapers).[12] Based on this data, we construct a measure of traditional newspaper consumption as the number of newspaper sales per capita.

For our comparison of social and more traditional media, we collected the number of total and refugee-related reports in German news media from Nexis UNI (previously LexisNexis). We use this to construct the weekly share of news reports about refugees. For further analysis, we obtained the full text of all refugee-related reports using the Lexis bulk data API, as well as all Facebook data from the pages of five major German newspapers (Welt, Frankfurter Allgemeine Zeitung (FAZ), Tageszeitung (TAZ), Süddeutsche Zeitung (SZ), and Bild).

We also include controls for the local prevalence of right-wing extremism. One such measure is the number of murders committed by neo-Nazis in each municipality from 1990 until 2016, which were collected by "Mut gegen rechte Gewalt" (Courage Against Right-Wing Violence). We complement this proxy for contemporary right-wing violence with data on the

[11] Broadband access is highly correlated with publicly available survey data on individuals' internet use from Eurostat; these data are only available on the state level (see Figure 1.10 in the online appendix).

[12] These data contain the number of print newspapers sold in each municipality with more than 3,000 inhabitants. Newspapers are listed if, in any given town, they (1) sell at least 50 copies and (2) have a market share of at least 1%. To have a similar sample size across specifications, we impute values for 1,120 towns for which news paper sales data are not available, based on a municipality's population, population density, AfD vote share, and county fixed effects. However, the results are almost equivalent without imputation (available upon request).

historic prevalence of anti-semitism collected by Voigtlander and Voth (2012).[13]

Finally, we obtain Google trends data on overall interest in the search terms "Brexit", "Trump", and "UEFA EM 2016" in Germany to proxy for distracting news events. Google scales the weekly number of searches for these terms on a scale from 0 to 100, where 100 marks the week with the highest search interest in the preceding 5 years. The time series plots in Figure 1.15 in the online appendix suggest these measures are sound approximations for attention paid to Brexit, the Trump election, and the UEFA European Championship (one of the most widely followed sports events in Germany).

1.3 Empirical Strategy and Main Results

1.3.1 Empirical Strategy

We begin to investigate the link between social media and anti-refugee incidents by estimating fixed effects panel regressions akin to a Bartik-type approach (Bartik, 1991). In particular, we use the interaction of local right-wing Facebook usage ($AfD\ Users/Pop_i$) and weekly refugee posts on the AfD Facebook page ($Refugee\ Posts_t$) to measure the differential change of hate crimes conditional on anti-refugee sentiment on social media. This empirical set-up creates variation by week and municipality, which we exploit in the following regression model:

$$\begin{aligned} Refugee\ attack_{it} = {} & \beta\ AfD\ Users/Pop_i \times Refugee\ Posts_t \\ & + \gamma\ Controls_i \times Refugee\ Posts_t \\ & + Week\ FE_t + Municipality\ FE_i + \epsilon_{it}, \end{aligned} \tag{1.1}$$

The dependent variable is a dummy for the incidence of a refugee attack in municipality i in week t. β measures the differential change in anti-refugee incidents conditional on Germany-wide posts about refugees on the AfD page—as a proxy of Germany-wide anti-refugee sentiment on social media—and right-wing social media users per capita. We control for a host of local characteristics interacted with the refugee post measure. Because we include many fixed effects and interaction terms, we estimate 1.1 using Ordinary Least Squares,

[13] From their dataset, we use the natural logarithm of one plus the number of deported Jews as well as one plus the number of letters written to "Der Stürmer", the antisemitic newspaper published by Nazi politician Julius Streicher. Towns with no information are coded as zero. We do not use scaled variables because the data from Voigtlander and Voth (2012) only cover a fraction of the municipalities in our sample.

which yields the linear probability model. Standard errors are clustered by municipality. We consider alternative specifications of the dependent variable and standard errors in robustness exercises.

This framework has three key features. First, it circumvents reverse causality, because refugee incidents in one town are unlikely to change anti-refugee sentiment in *all other* towns. Second, our measure of social media exposure is time-invariant and thus not the result of whether a municipality experiences refugee attacks in a given week.[14] Third, a full set of fixed effects controls for unobserved heterogeneity that affects all towns at the same time (such as salient news events), as well as time-invariant differences across towns (such as a history of anti-minority violence).

The main concern with estimating Equation (1.1) is that $AfD\ Users/Pop.$ may be correlated with other municipality characteristics that could explain differences in how local anti-refugee attacks co-vary with the salience of refugees online. In that case, we would not be capturing a pure social media "effect". For example, the share of AfD Facebook subscribers may partially pick up general right-wing attitudes, which could lead to more anti-refugee attacks in times of high refugee salience. This concern may also not be sufficiently addressed by controlling for interactions of observable municipality characteristics with the refugee salience measure.

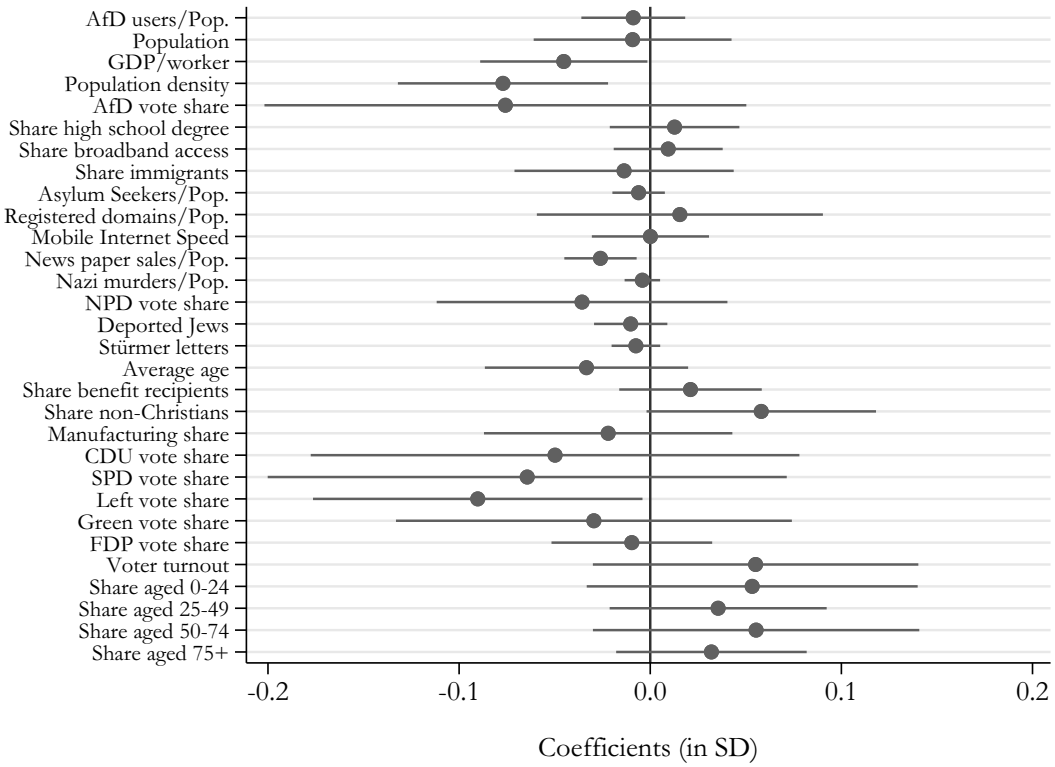
We therefore develop an identification strategy based on Facebook and internet outages. These disruptions induce plausibly exogenous variation in people's exposure to social media while leaving other local characteristics unchanged. The first part of this empirical strategy exploits the timing of major server problems at Facebook, which disrupt access to the platform. In the second part, we build on the insight that German internet infrastructure is trailing behind that of many other European Countries (e.g. Latvia) and the OECD average (see Times, 2017a; OECD, 2016). As a result, prolonged internet outages are relatively common. Because around 50% of worldwide Facebook users accessed the platform with their computers, many users are exposed to disruptions in internet access. In Germany, this share is likely to be even higher because of the relatively slow adoption of mobile internet.[15]

Local internet outages are widely dispersed geographically: Figure 1.11a visualizes the

[14] In the robustness section below, we alternatively measure local social media penetration before the start of the refugee crisis, at the cost of reducing the number of users for whom we have location data. This adjustment makes little difference for the results.

[15] Data on Facebook usage patterns reported on Statista.com and on mobile internet usage in Germany on (also on Statista.com) support this assessment.

Figure 1.3: Balancedness — Internet Outages and Local Characteristics



Notes: This figure plots the coefficients of the regression $\overline{Internet\ outages}_i = \alpha + \mathbf{X}'\beta + \epsilon_i$, where the dependent variable is the total number of internet outages in a municipality (based on our baseline definition) and $\mathbf{X}$ is a vector of local characteristics for which we plot the estimates. To make the magnitudes comparable, we standardize all variables to have a mean of zero and standard deviation of one. 95% confidence intervals are based on standard errors clustered by municipality.

distribution of disruptions per capita across Germany. The outages are also not particularly clustered in a particular time period (see Figure 1.11b). Crucially, the frequency of internet problems is uncorrelated with the share of the population on the AfD Facebook page. As such, internet disruptions provide exogenous variation that is not already captured by our variable on local Facebook usage. The number of reported internet problems is also uncorrelated with the total number of refugee attacks in a given municipality. In fact, regressing the frequency of internet outages on a host of municipality characteristics in Figure 1.3 suggests that they are largely uncorrelated with observable factors: the estimated coefficients are nearly all statistically indistinguishable from zero and quantitatively small. Taken together, our interpretation is that whether an internet outage occurs in a given town and week is as good as randomly assigned with regard to unobserved other factors that might drive hate crimes.

We analyze the effect of Facebook and internet outages in a flexible empirical framework. We begin by asking whether these outages reduce anti-refugee attacks, and whether they do so particularly in areas with a higher concentration of AfD Facebook users. We then study whether these disruptions also decrease our baseline correlation of local exposure to anti-refugee sentiment and hate crimes. More formally, the most saturated regressions have the following triple difference form:

$$\begin{aligned} Refugee\ Attack_{it} = {} & \beta\ AfD\ Users/Pop_i \times Refugee\ Posts_t \\ & + \lambda\ Outage_{it} \times AfD\ Users/Pop_i \times Refugee\ Posts_t \\ & + \delta_1\ Outage_{it} + \delta_2\ Outage_{it} \times Refugee\ Posts_t \\ & + \delta_3\ Outage_{it} \times AfD\ Users/Pop_i \\ & + \gamma_1\ Controls_i \times Refugee\ Posts_t \\ & + \gamma_2\ Controls_i \times Outage_{it} \\ & + Week\ FE_t + Municipality\ FE_i + \epsilon_{it}, \end{aligned} \tag{1.2}$$

For the Facebook outages, which only vary by week, we replace $Outage_{it}$ with $Outage_t$.[16] For the initial tests, we focus on the estimates for δ_1 and δ_3 while excluding the coefficients β, λ, δ_2, and γ_1. That is, we ask whether outages reduce anti-refugee incidents, and whether they reduce them more in areas with more AfD Facebook users. In the fully interacted regressions, the main coefficient of interest λ captures the correlation of anti-refugee attacks and local exposure to anti-refugee sentiment on social media, depending on whether an outage occurs. Put differently, we test whether outages break the correlation between real-life incidents and refugee salience, particularly for areas with high right-wing Facebook penetration. The vector $Controls_i \times Outage_{it}$ controls for the differential effect of outages based on observable characteristics, such as internet affinity.

The identifying assumption of this approach is that Facebook and internet outages only affect anti-refugee incidents through their effect on social media exposure. This assumption is plausible for Facebook outages. In the case of internet outages, for which we have variation at the municipality-week level, one may be worried about alternative online channels. We discuss these and other potential threats to identification in the next section.

[16]Note that, as a result, the estimates of δ_1 and δ_2 in Equation (1.2) are absorbed by the week fixed effects.

Exploiting variation in Facebook and internet outages also allow us to address the concern that towns with a stronger right-wing presence may show differential trends whenever the nationwide sentiment towards refugees changes. This is because these relatively short-lived outages are unlikely to affect the presence of deep-rooted right-wing attitudes in a municipality; absent online channels, the outages should thus not have an impact on real-life outcomes. The framework in Equation (1.2) further addresses reverse causality concerns. If we were merely capturing that local incidents drive posts on social media, Facebook and internet outages should not reduce the number of hate crimes. Instead, they should only reduce social media activity, keeping the number of anti-refugee incidents unchanged.

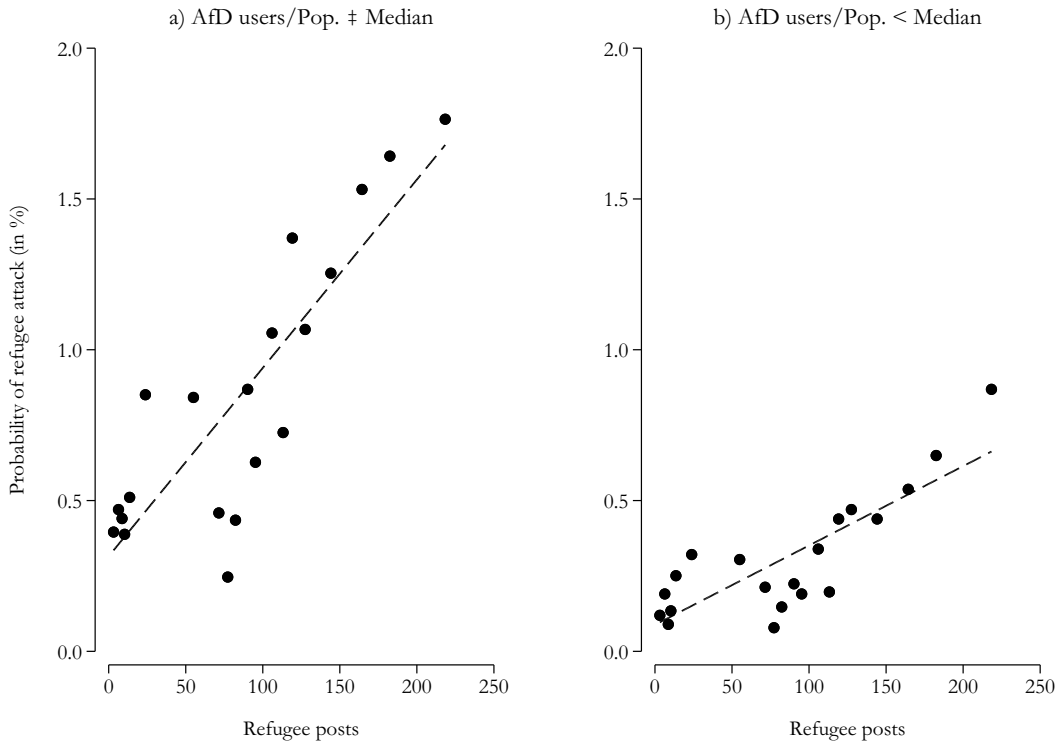
1.3.2 Panel Regression Results

We illustrate the intuition behind our regression framework in Figure 1.4. The figure shows a binned scatter plot of anti-refugee attacks and anti-refugee sentiment, split by the degree of exposure to right-wing social media. Higher refugee salience is associated with a higher probability of anti-refugee attacks in both sub-samples, but the positive slope is far more pronounced for towns with an above median AfD user to population ratio (Panel (a)). Our baseline regression coefficient picks up the difference in slopes between municipalities with high and low Facebook usage.

Table 1.2 presents the regression results from estimating Equation (1.1) with varying sets of control variables (interacted with refugee salience). The coefficient on the interaction of local Facebook usage and Germany-wide refugee posts is positive and highly significant in all specifications. Column 1 shows the panel regressions with the baseline control variables, which yields a coefficient 0.024 on the interaction term. This correlation does not appear to be driven by support for the AfD alone: the result holds although we control for the AfD vote share in the 2017 federal election. This highlights a distinction between our social media measure and general support for the party.

To get a sense of the magnitudes, consider as a case study the cities of Bochum and Hannover, which are about one standard deviation apart in the ratio of AfD users to population (in 1000s) (≈ 0.29). Holding average anti-refugee sentiment in our data constant (84 posts), this means a one standard deviation higher right-wing social media usage is associated with a 10% higher probability of an anti-refugee incident relative to the mean. Table 1.19 in the online appendix shows that this correlation is largely driven by cases of assault.

Figure 1.4: Exposure to Refugee Sentiment on Facebook and Hate Crimes



Notes: This figure plots the average number of anti-refugee attacks against our measure of anti-refugee sentiment for municipalities below and above the median of $AfD\ Users/Pop$. Refugee attacks are binned by 20 quantiles of refugee posts and residualized with respect to population.

In columns 2 through 6, we introduce a richer set of controls that accounts for local right-wing attitudes, general media exposure, more socio-economic factors, and the vote shares of all major parties in the 2017 election (see Table 1.10 for an overview of the control variables). In column 7, we add all interacted controls jointly. The inclusion of these covariates makes little difference to our main estimate. This is a first indication that the correlation between social media exposure and anti-refugee incidents is not driven by observable municipality differences unrelated to Facebook usage.

Table 1.2: Baseline Correlations — Facebook Posts and Hate Crime

		Additional interacted controls					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Baseline controls	Right Wing controls	Media controls	Socio-economic controls	2017 vote controls	Age structure controls	All controls
AfD users/Pop. × Refugee posts	0.024***	0.020**	0.023**	0.024**	0.021**	0.023**	0.016**
	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)
Observations	479,964	479,964	479,964	474,303	479,964	476,856	474,303
R-squared	0.082	0.083	0.082	0.083	0.083	0.083	0.084
Municipalities	4324	4324	4324	4273	4324	4296	4273
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Right-wing controls [4] × Posts		Yes					Yes
Media controls [4] × Posts			Yes				Yes
Socio-econ. controls [4] × Posts				Yes			Yes
Election controls [7] × Posts					Yes		Yes
Age controls [4] × Posts						Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). All control variables are interacted with the *Refugee posts* measure; see text for a description of the controls. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

1.3.3 Quasi-Experimental Evidence: Facebook and Internet Outages

To isolate the importance of social media, we next draw on internet and Facebook outages as sources of quasi-experimental variation. To count as a severe internet disruption, our baseline measure has to fulfill two criteria: (1) it has to last at least 24 hours, and (2) it has to affect a significant part of the population, i.e. be in the top quartile of reported internet disruptions

per capita, which vary by municipality and week (see section Section 1.2 for more details). This gives us 313 severe internet outages.[17]

Internet outages. Are local internet outages severe enough to decrease a municipality's exposure to social media? We investigate this question by using a sample of posts from the AfD Facebook page for which we know the users' locations.[18] Figure 1.5a plots the local number of posts against the intensity of local internet outages. Local Facebook activity falls with outage intensity and is close to 0 as soon as we observe more than 0.25 outage reports per 10,000 inhabitants. Figure 1.12 shows that we observe significantly fewer posts and comments on Facebook for municipalities that experience an internet disruption. These results lend credence to the idea that exposure to social media content is reduced in the affected municipalities and not compensated by users accessing Facebook with their mobile phones.

If internet outages indeed reduce local social media exposure, we would expect them to mediate the capacity of social media to propagate anti-refugee incidents. As described in Section 1.3.1, we test this hypothesis by interacting the main terms of interest $AfD\ Users/Pop_i \times Refugee\ Posts_t$ with $Internet\ Problems_{it}$, our dummy for severe internet disruptions. We graphically illustrate the results in Figure 1.5b. The binned scatter plot is almost identical to Figure 1.4, except that we plot a separate slope for municipalities that experience an internet outage. This reveals a striking pattern: while anti-refugee attacks increase with anti-refugee posts, this relationship disappears in municipalities that experience an internet outage. This holds true for municipalities with high and low Facebook usage.

Figure 1.5b implies that internet outages have a substantial attenuating effect. Consider the pattern in panel (a). Without outages, there is a strong correlation of refugee posts and attacks. During outages, the correlation is essentially zero. This means that the outage effect is larger than the baseline estimate for $AfD\ Users/Pop. \times Refugee\ posts$, which is given by the slope difference of the dotted lines in panels (a) and (b). We interpret this as evidence that cutting of users from social media completely has large effects.

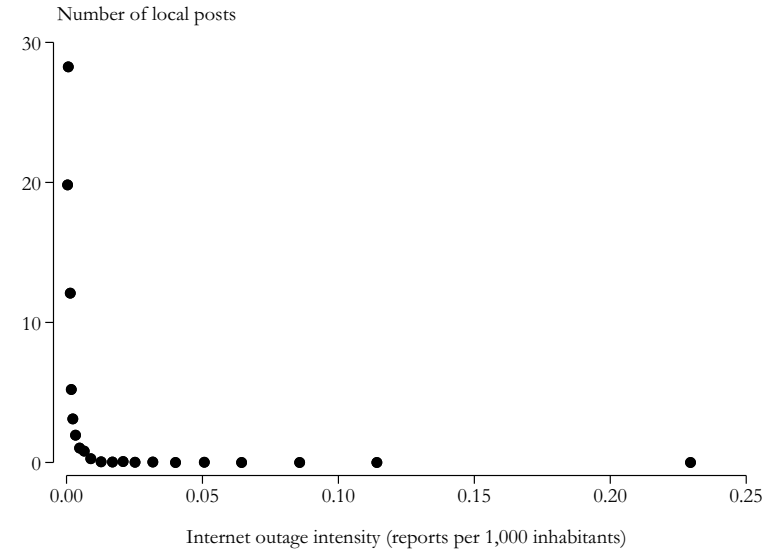
We next estimate versions of Equation (1.2) and report the regression results in Table 1.3. Column 1 shows that internet outages reduce anti-refugee violence. The coefficient of -0.003

[17] In the online appendix, we show our results are robust to alternative definitions. We also exploit the eight major Facebook outages, which only vary by week. We discuss the results and their interpretation in turn.

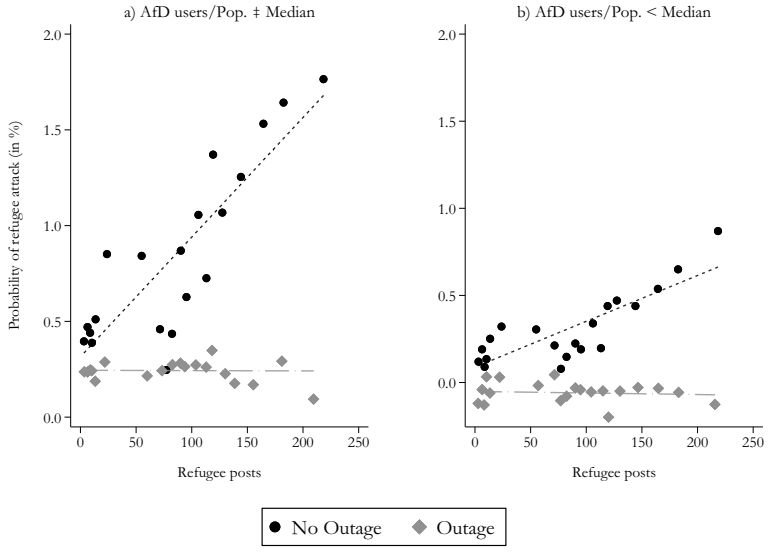
[18] These posts and comments are a sub-sample by users who publicly disclosed their location in their Facebook profiles.

Figure 1.5: Quasi-Experimental Results from Internet Outages

(a) Internet Outages Reduce Local Facebook Activity



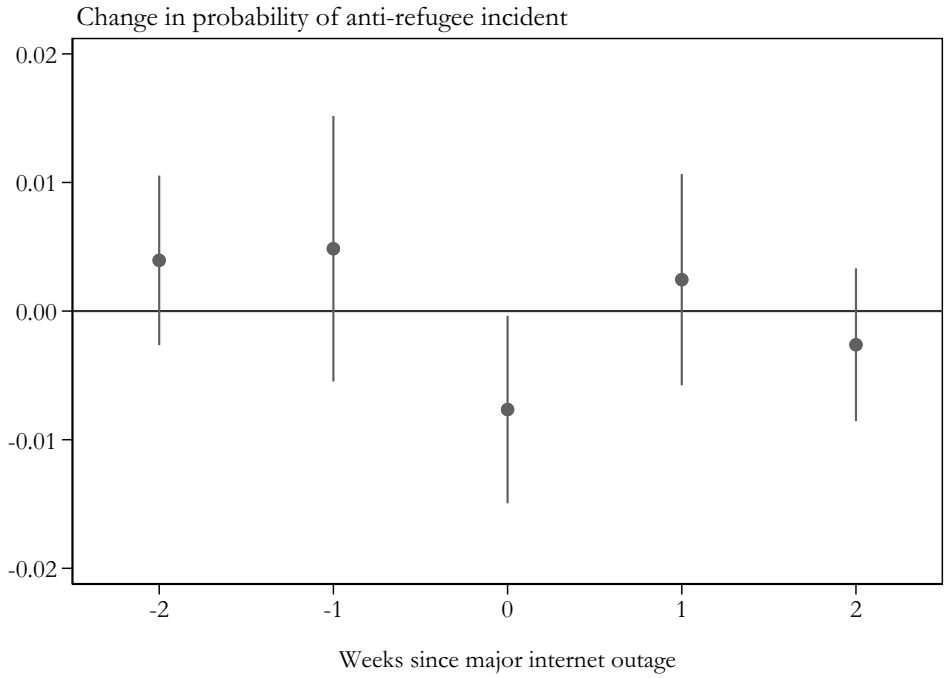
(b) Internet Outages Reduce Local Anti-Refugee Incidents



Notes: Panel (a) shows a binned scatter plot of local posts on the AfD Facebook page as a function of the reports on internet outages in a given week. Panel (b) plots the average number of anti-refugee attacks against our measure of anti-refugee sentiment for municipalities above and below the median of $AfD\ Users/Pop$. Refugee attacks are binned by 20 quantiles of refugee posts. We additionally split towns by whether they experience an internet outage in a given week (gray squares). The number of anti-refugee attacks is residualized with respect to population; hence, the number of attacks can be slightly below 0 in some bins.

implies that, during such outages, the probability of a refugee attack is 53% lower relative to the dependent variable mean (≈ 0.006). In Figure 1.6, we investigate the timing of this drop in incidents. Because the outages are relatively rare in the municipality-week panel, the estimates are necessarily noisy. Nonetheless, we can see a reduction in anti-refugee incidents that is sharply concentrated in the week of the internet outage.

Figure 1.6: Internet Outage Event Study



Notes: This figure plots estimates the estimates for δ from the event study regression $Attacks_{it} = \sum_{t=-2}^{2} \delta_{w=t} Outage_{it} + Fixed\ Effects + \epsilon_{it}$, where $Outage$ refers to internet outages in municipality i in week t. 95% confidence intervals are based on standard errors clustered by municipality.

Column 2 in Table 1.3 implies that this effect is driven by periods of high sentiment; it may also be driven by areas with many AfD Facebook users (column 3) but the coefficient is not statistically significant. In columns 4 through 6, we estimate the full triple-difference model. Here, we estimate the effect of outages in areas with high social media use at times of high anti-refugee sentiment. The estimates suggest that internet problems reduce social media's impact on anti-refugee violence. While the coefficient of refugee posts and social media exposure is similar to our baseline correlations, the triple interaction term with internet outages is negative and statistically significant in all three specifications. Quantitatively, internet outages appear to mitigate the entire effect of social media. In line with the graphical evidence in Figure 1.5b, we find that the triple interaction coefficient is larger than the baseline

coefficient. Put differently, for a given level of anti-refugee sentiment, there are fewer attacks in municipalities with high Facebook usage during an internet outage than in municipalities with low Facebook usage *without* an outage.

Table 1.3: Local Internet Outages and Social Media Transmission

	(1)	(2)	(3)	(4)	(5)	(6)
Baseline Interaction						
AfD users/Pop. × Refugee posts				0.024***	0.016**	0.016**
				(0.009)	(0.008)	(0.008)
AfD users/Pop. × Posts × Outage				-0.181***	-0.184***	-0.172***
				(0.058)	(0.058)	(0.057)
Outage Interaction						
Outage	-0.003***	-0.000	-0.003**	-0.001	-0.002	-0.007
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.008)
Refugee posts × Outage		-0.005***		-0.000	0.001	0.000
		(0.001)		(0.002)	(0.002)	(0.002)
AfD users/Pop. × Outage			-2.685	4.441	4.455	4.391
			(3.464)	(4.384)	(4.054)	(4.058)
Internet Usage Interaction						
Share broadband access × Outage						-0.000
						(0.000)
Internet domains/Pop. × Outage						0.021*
						(0.012)
Mobile Broadband Speed × Outage						0.000
						(0.000)
Observations	479,964	479,964	479,964	479,964	474,303	474,303
R-squared	0.082	0.082	0.082	0.082	0.084	0.084
Municipalities	4324	4324	4324	4324	4273	4273
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes	Yes	Yes	Yes
All other controls [22] × Posts					Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). Internet outages are defined as municipality-weeks that are in the top quartile of the ratio of reported internet outages to population. The coefficient of "Refugee posts × Outage" is multiplied by 100 for readability. Columns 1-4 include the baseline controls. Columns 5 and 6 include all controls as in column 7 of table 1.2, interacted with *Refugee posts* (unreported). Column 6 further adds the interaction of broadband access and internet domains/pop. with local internet outages. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Could it be that the effect of internet outages is merely coincidental? As an alternative way of assessing statistical significance, we perform a randomization test. Instead of the actual internet disruptions, we randomly define 313 municipality-week pairs as placebo outages. We then estimate the same regression using 500 different sets of placebo outages. This allows

us to evaluate the probability of finding a statistically significant coefficient in our dataset. Using this procedure, we find that more than 99% of the placebo triple interaction coefficients exhibit a lower t-statistic than our estimate. Our findings are thus unlikely to be purely coincidental. We show the full distribution of t-statistics from this randomization test in Figure 1.14a in the online appendix.

The identifying assumption for internet outages in our framework is that they only have an effect on anti-refugee hate crime through the reduced exposure to social media. Could it be that we observe reduced hate crimes because users are cut off from the internet generally, not from social media in particular? Two pieces of evidence support the idea that we capture a social media channel.

First, when we include interactions of internet disruptions with measures of internet usage (broadband access, per capita internet domains, mobile internet access), our main coefficient is unaffected (see column 6 in Table 1.3). The coefficients of the internet usage interactions are generally statistically insignificant or have the opposite of the expected sign. This is at least some indication that we are not merely capturing general internet usage. It also suggests that our findings are unlikely to capture that people are busy fixing internet access problems. If we were merely capturing such displacement effects, one would expect it to more strongly affect people in areas with high internet usage, which does not seem to be the case in the data. Second, after including the other interaction terms in columns 4 through 6, the coefficient on internet outages is no longer statistically significant. This result also supports the idea that internet outages reduce hate crime by limiting access to social media.

Another concern could be that hate crimes are less likely to be reported during internet outages. We believe this is unlikely to explain our findings because we analyze incidents that happened years in the past. While internet outages might hamper the flow of information, it seems highly unlikely that incidents such as assault or property damages are *never* reported due to a temporary internet disruption. As further evidence, we limit our analysis to official reports by the police or the German parliament, for which social media reporting is an unlikely concern. This yields similar results (see column 1 of Table 1.15).

We also run a number of tests to rule out that our Germany-wide measure of refugee posts is affected by local internet outages. As stated above, this appears unlikely because we focus on *local* disruptions to the internet; Table 1.14 in the online appendix shows that the total number of internet outages in a given week is uncorrelated with the total number of Facebook posts. The outage results are also robust to using a leave-one-out measure of

refugee posts (column 2), Germany-wide posts in the previous week (column 3), and an alternative measure based on Google search intensity for the word refugee (*Flüchtling*) in column 4. The implied magnitudes are almost equivalent.[19] This suggests that the outage effect is driven by exposure rather than the production of anti-refugee content. In Table 1.17, we show additional robustness checks for alternative transformations of the dependent variable. The findings remain robust throughout. Table 1.18 shows that the results also hold using alternative definitions of the outage dummy.

Facebook outages. As further evidence for the social media transmission mechanism, we use eight major Germany-wide Facebook outages as a source of exogenous variation specific to social media access. Table 1.13 outlines the details of each of the eight outages and links to relevant press reports. By definition, these outages are Facebook-specific and therefore do not affect other potential channels of online transmission.

Table 1.14 in the online appendix shows that these outages are large enough to disrupt weekly activity on right-wing social media. Column 1 and 2 show that, during weeks with Facebook outages, there are on average 11% fewer new total posts and 24% fewer posts about refugees on the AfD page.[20] There is no evidence of such an effect in the week before. Column 5 shows that Facebook outages are also uncorrelated with the total number of weekly internet disruptions ($t = -0.41$).

We next present the results of interacting Facebook disruptions analogous to the internet outages in Table 1.4. The results again reveal a clear pattern. The coefficient of -0.001 in column 1 shows that the probability of an anti-refugee incident is around 18% lower in weeks with major Facebook outages (relative to the unconditional probability of an attack). Figure 1.13 suggests that the timing of this effect is concentrated in the week of the Facebook outage, without significant effects in the week before or after the outage. Because we solely rely on the weekly variation from the few major Facebook outages, the estimates are noisier than those for internet outages. Column 2 shows that, intuitively, this effect is also larger in areas with many users on the AfD Facebook page. The coefficient of 2.222 suggests that Facebook outages reduce the probability of a hate crime by 12% more for a one standard

[19] To see this, consider the effect implied by dividing the triple interaction coefficients by the standard deviation of these salience metrics. This suggests that internet outages have a mediating effect of 9.6, 10.5, and 11.0 for the AfD posts about refugees, the leave-one-out measure, and Google trends, respectively.

[20] The average number of refugee posts in the time series is around 84. The coefficient estimate of 19.880 implies an effect of Facebook outages on posts of $-19.880/84 \approx 0.24$ relative to the mean.

deviation increase in $AfD\ users\ /Pop.$.[21] This is additional evidence that social media *per se* might affect hate crimes.

Next, we introduce the triple interaction of Facebook outages with social media usage and our refugee salience measure. The triple interaction is negative and statistically significant in all three specifications in columns 3 through 5. Quantitatively, we find that Facebook disruptions fully undo the baseline correlation of refugee attacks and exposure to social media sentiment. For example, consider that the coefficient of $AfD\,users/Pop.$ and $Refugee\,Posts$ is 0.027 in column 4 but -0.04 on the triple interaction. This implies that, in weeks of major Facebook outages, heightened refugee sentiment is not associated with a differential increase of anti-refugee attacks in municipalities with higher Facebook usage.

It is worth noting that we would expect the Facebook outage coefficients to differ in magnitude from the internet outage coefficients. This is because Facebook outages eliminate the differential exposure *between* areas with high and low social media usage to anti-refugee posts. In contrast, internet outages further exploit variation *within* municipalities. Because within-municipality variation induced by internet outages appears to matter more in our setting, we find smaller coefficients for Facebook outages.

We again perform a randomization test to assess the statistical significance of the Facebook outage results. We randomly assign placebo Facebook outages to eight weeks in our data, excluding the weeks in which we identified Facebook outages. We then estimate the same regression using 500 different sets of placebo outages. Using this procedure, we find that 92% of the placebo triple interaction coefficients exhibit smaller t-statistics. We show the full distribution of t-statistics from this randomization test in Figure 1.14b in the online appendix. This confirms that our findings are unlikely to be a matter of coincidence.

Taken together, the evidence here suggests that the relationship of anti-refugee sentiments online and hate crimes is attenuated by Facebook and internet outages. These results are most consistent with a causal propagation effect of social media.

In the online appendix, we conduct additional robustness exercises for our outage results. In Table 1.16, we show a range of different standard errors. We also assess our results' robustness to different transformations of the refugee attack variable and estimation methods in Table 1.17. Our results are similar when we use the number of attacks, log(1+refugee

[21] In unreported results, we also find that the interaction of Facebook outages with refugee posts has a statistically significant negative coefficient.

Table 1.4: Facebook Outages and Social Media Transmission

	(1)	(2)	(3)	(4)	(5)	(6)
Baseline Interaction						
AfD users/Pop. × Refugee posts			0.027***	0.027***	0.021**	0.021**
			(0.010)	(0.010)	(0.009)	(0.009)
AfD users/Pop. × Posts × Outage			-0.040*	-0.040*	-0.046**	-0.046**
			(0.021)	(0.021)	(0.022)	(0.022)
Additional Outage Coeffcients						
Outage	-0.001***					
	(0.000)					
AfD users/Pop. × Outage		-2.222*	1.164	1.164	1.367	3.230
		(1.273)	(1.833)	(1.833)	(1.862)	(1.969)
Observations	479,964	479,964	479,964	479,964	474,303	474,303
R-squared	0.079	0.082	0.082	0.082	0.084	0.084
Municipalities	4324	4324	4324	4324	4273	4273
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE		Yes	Yes	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes	Yes	Yes	Yes
All other controls [22] × Posts					Yes	Yes
All controls [30] × Outages						Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). Facebook outages refer to weeks in which Facebook experienced considerable disruptions; see the online appendix for more details on how these are defined. Note that the other interaction terms *Outage*, $Refugee\ posts$ and $Outage \times Refugee\ posts$ are absorbed by the week fixed effects in columns 3-5. Columns 1-3 include the baseline controls. Columns 4 and 5 include all controls as in column 7 of table 1.2, interacted with *Refugee posts*. Column 5 adds the interaction of these control variables with Facebook outages. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

attacks) or the ratio of refugee attacks to asylum seekers as dependent variable. In all cases, the estimated coefficients are highly statistically significant.

1.3.4 Additional Results

Other Posts on the AfD Facebook Page: If the channel we uncover is indeed specific to refugees, we would expect a weaker correlation between refugee attacks and posts about other topics on the AfD Facebook page. We test this hypothesis in Table 1.20, where we plot the baseline estimation with refugee posts in column 1 for convenience. We also report coefficients for standardized post measures (with a mean of zero and standard deviation of one) in square brackets to compare coefficient sizes across the different posts. Next, we estimate Equation (1.1) using all posts except those containing the word *refugee* ("Flüchtling") in column 2. The estimate is statistically indistinguishable from zero. We also repeat our baseline test using posts containing the words "Muslim", "Islam", or "EU"—the latter is motivated by the AfD's long-standing criticism of the European Union. For all these terms, we find no significant relationship between the number of posts and the number of attacks; all estimated coefficients are considerably smaller in standardized terms compared to the baseline measure. This shows the specificity of our refugee measure: the correlation we capture does not appear to be an artifact of general anti-minority sentiment, but rather a predictable result of increased animosities towards refugees on social media in particular weeks.

Intensive Margin of Facebook Usage: If social media works as the propagating mechanism for hate speech, we would also expect its effect to increase with how frequently users interact with the AfD Facebook page. We explore this issue empirically in Table 1.21, where we interact our main interaction term with the total number of local posts on the AfD wall and the number of comments and likes on AfD posts, all scaled over the number of AfD users in a municipality.[22] These measures of usage intensity are not systematically correlated with local Facebook penetration, city size, or population density. As such, they create additional variation in social media engagement across towns.

[22]Note that we can only construct these measures on the intensive margin of municipalities where we can identify at least one AfD user. Our baseline results also hold in this sub-sample, which we show in Table 1.26 in the online appendix.

The results suggest that local engagement on Facebook matters: all three triple interaction terms are positive and statistically significant. Consistent with the hypothesis that social media enables hateful sentiment to spread, a higher reach per AfD user increases the correlation of social media exposure with hate crimes. These interactions work on top of our baseline interaction term, which remains similar in magnitude and highly statistically significant throughout. The smallest coefficient on the triple interaction term of 0.001 in column 3 implies that a one standard deviation increase in likes per user (around 12) increases the baseline coefficient by 25%.[23]

Distracting News Events: As an additional piece of analysis, we investigate the role of news shocks on the transmission of online hate speech to real-world actions. We build on the evidence in Durante and Zhuravskaya (2018), who show that the Israeli army is more likely to strike against Palestinian targets when US media outlets are distracted by other news events. In our case, we hypothesize that other important news events might distract people from the topic of refugees. This is somewhat analogous to Facebook outages in that we exploit additional exogenous weekly variation: if major news events act as a distraction, they should reduce the correlation of exposure to refugee salience with hate crimes.

To measure these news shocks, we obtain Google Trends data on weekly search interest on the terms "Brexit", "Trump", and "UEFA Euro 2016'. Figure 1.15 shows that these spike around the respective events. In Table 1.22, we show that they are indeed associated with a crowding out of refugee salience: the share of posts about refugees is markedly lower during these key events. As an example, the spike in search interest for Brexit (100 on the Google search index) is associated with an almost 30% drop in the share of refugee posts (relative to the mean).

We next investigate whether, as a result, refugee salience has a weaker link with hate crimes in the weeks these major events attracted particular news attention. If this is the case, we would expect that these events *decrease* the correlation of social media transmission with refugee attacks. As before, we implement this by including the Google trends measures as a further interaction in our panel regressions.

Table 1.5 plots the results. For each of the events in columns 1 to 3, we find a significant

[23]To see this, consider that the total implied estimate including interaction is calculated as $0.001 \times 12 \approx 0.012$, which is about 25% than the baseline coefficient of 0.049.

negative coefficient on the number of anti-refugee incidents for the triple interaction with distracting news. The negative sign of the coefficient indicates that, during weeks of major news events, changes in anti-refugee incidents correlate less with heightened refugee salience. As the salience of other events crowds that of refugees, there are smaller increases of hate crimes in municipalities with more AfD social media users.

Table 1.5: News Shock Salience and Hate Crime Propagation

	(1) Brexit	(2) Trump	(3) UEFA EM 2016
AfD users/Pop. × Refugee posts	0.071***	0.096***	0.067***
	(0.018)	(0.022)	(0.017)
AfD users/Pop. × Posts × News shock	-0.019**	-0.009***	-0.002**
	(0.008)	(0.003)	(0.001)
Observations	495,726	495,726	495,726
R-squared	0.078	0.079	0.078
Municipalities	4466	4466	4466
Municipality FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). The news shocks refer to the Google searches as indicated in the text. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01%, 0.05%, and 0.1% levels, respectively.

1.3.5 Differences Between Social Media And Traditional Media

How does social media differ from traditional media? And could such differences partially explain our results? Existing work has highlighted the ability of users to self-select and interact on social media (e.g. Schmidt et al., 2017). In the following, we highlight three aspects of far-right social media in Germany that may make it a particularly effective transmission mechanism for anti-refugee sentiment compared to mainstream news sources.

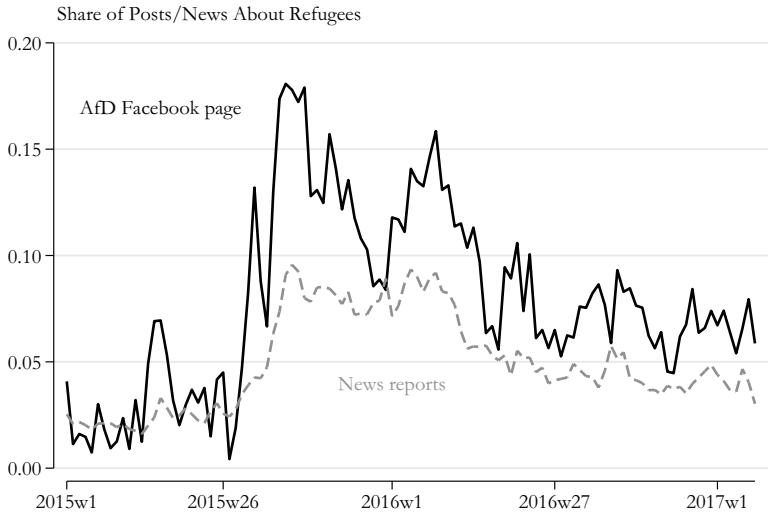
First, Figure 1.7a shows that the share of content about refugees is consistently higher on the AfD's Facebook page compared to traditional news outlets in the Nexis data. The share of refugee mentions on Facebook is also far more volatile and spikes coincide more clearly with salient news events like Merkel's "Wir schaffen das" speech or the Cologne New Year's Eve incidents. In both of these examples, the share of refugee posts on right-wing social media is nearly 100% higher than the share of news stories on refugees, which is consistent with the idea that the topics discussed on Facebook are considerably narrower than in traditional media.

In Figure 1.16a in the online appendix, we show that this also holds true in a like-for-like comparison of the share of refugee posts on the AfD's Facebook page relative to the Facebook pages of five major German news outlets. AfD users post twice as much about refugees compared to the next-ranked newspaper. This suggests that the narrowness of content is unlikely to be explained only be the editorial constraints (e.g. space limits in newspapers) of traditional media outlets. Instead, self-selection of like-minded people into the AfD Facebook page likely also play a role. Combined with the interactive nature of social media, this result points towards an anti-refugee group dynamic on the AfD's Facebook page.

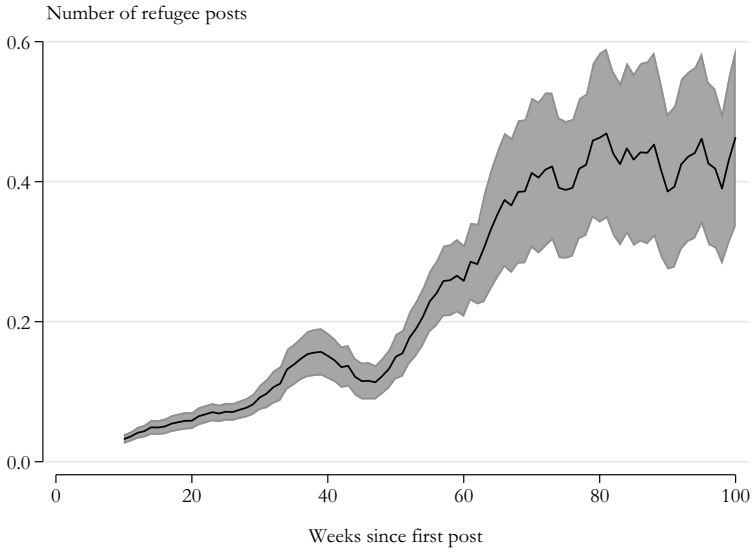
Second, as argued by Sunstein (2017), self-selection of like-minded people can lead to the expression of more extreme viewpoints. To shed light on this hypothesis empirically, we compare the full text of news reports about refugees with posts on the AfD Facebook page. Existing reports on far-right hate speech on social media highlight three characteristics as typical (see for example Dinar et al., 2016; Kreißel et al., 2018; Ott and Gür-Seker, 2019): (1) a belief to speak for the "true will" of the people, i.e. the in-group (citizens) compared to the out-group (refugees); (2) an opposition to "elites", in particular politicians and the media, who supposedly mislead or betray the people in an undemocratic way; and (3) a legitimization of discrimination against refugees by highlighting crimes by refugees, an alleged incompatibility of cultural differences, and negative repercussions for vulnerable "locals" (e.g.

Figure 1.7: Highlighting Social Media Echo Chambers

(a) Share of Refugee Post over Time



(b) Individual Posting Behavior, by Length of Exposure



Notes: Panel (a) plots the share of posts/reports about refugees on the AfD Facebook page and major German news outlets from Nexis. Panel (b) plots the 10-week moving average of the number of refugee posts per person as a function of a user's time spent on the AfD Facebook page, proxied by the time since the first post. The shaded area indicates 95% confidence intervals.

women, children or pensions).

We find evidence for all three of these features of right-wing hate speech on the AfD's Facebook page. Our approach is to investigate which words occur with a higher probability in posts on the AfD page relative to news reports in the Lexis corpus.[24] We filter words using the word stems of the German terms for people, elite, democratic, press, crime, foreign, culture, refugee, betrayal, and several vulnerable groups (pensioners, children, women, homeless).

The results of this exercise in Table 1.6 reveal a clear pattern (see also Table 1.23 in the online appendix). As one example, the term "Volksbetrug" (betrayal of the people) is 1715 times more likely to appear on the AfD page than in traditional news outlets. Criticism of "elites" and the media are also far more frequent. Another main difference is how often crimes by refugees are discussed, based on the use of loaded terms like "Flüchtlingskriminalität" (refugee crime). We see expressed fears about "Fremdkulturen" (foreign cultures) and "Burkafrauen" (burka women). This analysis clearly shows that far-right ideas that have widely been interpreted as hate speech are far more pervasive on the AfD page than in traditional media reports.

We find similar results using a text analysis approach using machine learning. In particular, we train a L1 regularized logistic regression model classifier that predicts whether a text comes from the AfD Facebook page or a traditional media outlet. The classifier thereby identifies the words with the highest predictive ability for posts on the AfD Facebook page. Figure 1.17 shows a word cloud of the 100 words that best separate social media from traditional media content, based on the model with the highest cross-validated out-of-sample F1 scores.[25] The size of the words represents the magnitude of the coefficients as a measure of variable importance. Consistent with the findings in Table 1.6, critiques of establishment parties and the economic or social costs of refugees are among the words that most uniquely identify posts on the AfD page.

Third, we investigate how individuals' posting behavior varies with the length of exposure to far-right social media content. We construct a balanced panel of users' activity on the AfD's Facebook page. In Figure 1.7b, we show users' average number of posts about refugees since their first post on the page. To avoid that a changing sample composition drives our

[24]We calculate word probabilities for each corpus by dividing the number of times a word is mentioned ($Word_i$) by the total number of words in the corpus ($\sum Words_i$), e.g. $P(Word_i^{News}) = Word_i^{News} / \sum Words_i^{News}$. The relative probability is the ratios between the two calculated the two probabilities, i.e. $P(Word_i^{Facebook})/P(Word_i^{News})$.

[25]Note that the model was run in German and the words translated by the authors afterwards. For more details on the machine learning model, see the notes to Figure 1.17.

Table 1.6: Relative Word Frequencies on the AfD Facebook Page

Rank	Word	Translation	Relativ prob.
Panel A: Flücht (refugee)			
1	Flüchtlingsenklaven	refugee enclave	780
2	Flüchtlingslüge	refugee lie	693
3	Flüchtlingsirrsinn	refugee insanity	650
4	Flüchtlingsmafia	refugee mafia	520
5	Flüchtlingsbefürworter	refugee supporter	520
Panel B: Krimi (crime)			
1	Regierungskriminalität	goverment crime	1300
2	Diskriminierungsgesetze	anti-discrimination laws	520
3	Schwerstkriminellen	dangerous criminals	260
4	Fluechtlingskriminalität	refugee crimes	260
5	Kriminalitätssteigerung	increase in crime	260
Panel C: Presse (media)			
1	Freie Presse	free press	390
2	Propagandapresse	propaganda press	260
3	Presseempfang	press meeting	260
4	Meinungspresse	opinionated media	260
5	Nazipresse	nazi media	260
Panel D: Volk (people)			
1	Volksbetrug	betrayal of the people	1715
2	volksfeindlich	hostile to the people	780
3	volksverdummenden	brainwashing the people	520
4	Volksverhetzungsparagraphen	law against incitement	520
5	Volksprotesten	protest by the people	260
Panel E: Verrat (betrayal)			
1	Volksverrats	betrayal of the people	130
2	Vaterlandsverrat	betrayal of the fatherland	43
3	Volksverrat	betrayal of the people	43
4	Hochverrat	high treason	36
5	verratenen	betrayed	32

Notes: This table plots the relative probability of words mentioned on the AfD Facebook page compared to reports by major German news outlets on Nexis. We report the results by groups of word stems identified as likely to reflecting right-wing hate speech on social media by previous work in Dinar et al. (2016).

results, we restrict the analysis to the approximately 60% of users who first interacted with the AfD page before June 2015 and thus have been active on it for at least 100 weeks. The results are similar without this restriction.

The frequency of refugee posts strongly increases with users' duration on Facebook: within the first year, the average user on the AfD page goes from close to zero to posting at least once about refugees every 2 weeks.[26] This result suggests that the AfD page does not merely attract already active Facebook users with right-wing views, but may increase the willingness of people to express anti-refugee views over time.

This analysis also highlights an important distinction compared to existing research on media and violence. Yanagizawa-Drott (2014) Adena et al. (2015), and DellaVigna et al. (2014) all investigate the effect of nationalistic propaganda in settings of high ethnic tensions. In our setting, there is no nationalistic anti-minority propaganda in traditional media outlets. Rather, we find that social media provides an alternative forum to exchange and spread extreme rhetoric and viewpoints for the fringe elements of society.

1.3.6 Mechanisms

In theory, multiple mechanisms could be consistent with social media playing a propagating role in real-life hate crimes. We discuss four mechanisms: information exchange, persuasion, collective action, and local spillovers. We provide suggestive evidence that collective action and local spillovers likely play a role in our setting.

First, social media might facilitate the exchange of information. In our setting, relevant information for potential perpetrators could, for example, include the locations of refugee homes and meeting points for demonstrations. We analyze the content of the refugee posts on the AfD Facebook to identify any post that might contain location information. To do so, we tag posts that either contain a zip code, mention the word "straße" (street), "weg" (path), "Flüchtlingsheim", "Asylantenheim", "Flüchtlingsunterkunft" (all three translate to refugee home) or refer to a name of a German town or village.[27] We then manually check the content of tagged posts. This analysis suggests that while some locations like Berlin and Cologne are frequently mentioned in the posts as references to politicians and crimes committed by refugees, we find no mention about specific local information. We found no instance of zip

[26] The same holds true for the total number of posts (see Figure 1.16b in the Online Appendix).

[27] We base the search on a comprehensive list of 2,061 German towns and 11,000 municipalities from the German statistical office, which covers villages with as little as 20 inhabitants.

codes or exact addresses. It hence appears unlikely that this channel is the primary driver behind our findings.

A second mechanism could be a persuasion channel, implying that social media persuades potential perpetrators that refugees may be dangerous or undeserving, which may then push some people over the edge. We believe that the timing in our setting makes this channel unlikely. In contrast to other work in Müller and Schwarz (2018b) and Bursztyn et al. (b), we focus entirely on high-frequency variation in social media posts and refugee violence. To the extent that social media changes people's attitudes, this is unlikely to happen in a single week and revert back after anti-refugee salience has subsided. This is particularly true for the results on Facebook and internet outages: it seems unlikely that being cut off from social media during such disruptions reduces hate crimes because potential perpetrators become less xenophobic for a single week.

Third, social media could motivate collective action. Existing evidence in Enikolopov et al. (2020) and Manacorda and Tesei suggests that social media and mobile internet increase the incidence of protests. In our setting, users could coordinate to carry out hate crimes or learn about others' willingness to carry them out via social media. To investigate this, we rerun the panel regressions in Equation (1.1) but limit refugee attacks to those undertaken by multiple perpetrators.[28] In line with the collective action hypothesis, Table 1.7 suggests that our panel regression results are predominantly accounted for by cases with four or more perpetrators. We find no relationship for incidents with fewer than 4 perpetrators. Within the sub-sample where we can identify the number of perpetrators, these attacks account for a similar number of total incidents compared to the cases with more than 4 perpetrators. Hence, this finding is unlikely to be the result of limited statistical power.

Fourth, and somewhat relatedly, it could be that social media enables local spillovers, e.g. through "copy-cat" incidents. This mechanism suggests that potential perpetrators may use social media to learn about other attacks taking place, which could inspire them to carry out additional hate crimes. Because friendship networks on social media are clustered geographically (Bailey et al., 2018), this should be particularly pronounced for attacks happening nearby. We thus again rerun the panel regressions in Equation (1.1) but now include a dummy variable if neighboring municipalities experience an attack in a given week.[29]

[28] We were able to hand-code the number of perpetrators for 28% of the hate crimes.

[29] This is akin to the common correlated effects (CCE) estimator proposed by Pesaran (2006) to hold common shocks constant.

Table 1.7: Mechanism — Anti-Refugee Incidents, by Number of Perpetrators

	(1) Known perp. sample	(2) 1 perp.	(3) <4 perp.	(4) ≥4 perp.
AfD users/Pop. × Refugee posts	0.010**	0.003	0.004	0.007**
	(0.005)	(0.002)	(0.003)	(0.003)
Observations	479,964	479,964	479,964	479,964
R-squared	0.081	0.037	0.046	0.055
Municipalities	4,324	4,324	4,324	4,324
Share of attacks	1	0.245	0.494	0.534
Mean of DV	0.002	0.000	0.001	0.001
Municipality FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1), where we vary the definition of the dependent variable based on the number of perpetrators. All control variables are interacted with the *Refugee posts* measure. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Table 1.24 suggests that hate crimes happening in the same week nearby are associated with more anti-refugee incidents. This correlation strongly interacts with the popularity of right-wing social media, particularly when anti-refugee sentiment is elevated. In other words, having an attack in a neighbouring municipality is associated with a stronger correlation of exposure to right-wing social media and the probability of an anti-refugee incident.[30]

Overall, our results appear to be most consistent with the idea that short-run bursts in anti-refugee sentiment on social media can translate into real-life hate crimes by enabling coordination online, both through group actions and local spillovers.

1.3.7 How Many Refugee Attacks Are Caused By Online Hate Speech?

We conduct a back-of-the-envelope calculation of how many attacks against refugees would have taken place with lower anti-refugee sentiment on right-wing social media. Given that we rely on high-frequency variation, this question is difficult to address. As our estimates are likely to pick up two separate facets of exposure to social media.

On one hand, it could be that exposure to anti-refugee sentiment on social media merely affects the exact timing when refugee attacks occur without changing their total number. On the other hand, the time series of hate crimes and refugee posts on social media in Figure 1.2 exhibits prolonged overall increases in the number of anti-refugee incidents with the onset of the refugee crisis. These increases are not easy to explain if anti-refugee sentiment exclusively affects the timing of incidents. In our empirical setting, we cannot distinguish between these possibilities.

Despite this important caveat, we still believe it is instructive to assume social media indeed increases the number of hate crimes to illustrate the magnitudes of the results. We calculate the predicted number of attacks, based on the coefficient estimate of 0.024 from a regression with the baseline control variables (see column 1 in Table 1.2). Multiplying this coefficient with $AfD\ users/Pop.$ and $Refugee\ posts$ gives us the estimated effect on anti-refugee attacks. We sum over all observations to get the total predicted number of anti-refugee attacks as a result of social media. This calculation implies that in absence of social media transmission on social media would result in 289 (10%) fewer anti-refugee incidents.

[30]Note that, although they are suggestive, we do not interpret these estimates as causal "peer effects", because we cannot distinguish them from common shocks (see Manski, 1993).

1.4 Conclusion

Social media has become a powerful tool for sharing and disseminating information. In this paper, we investigate whether social media can play a role in propagating violent hate crimes. Our findings suggest that social media has not only become a fertile soil for the spread of hateful ideas but also motivates real-life action. By combining detailed local data on Facebook usage with user-generated content, we can shed light on the link between online posts and anti-refugee incidents in Germany. Plausibly exogenous variation in disruptions to users' Facebook or internet access supports the view that some of the correlations we document reflect a causal effect.

Existing research shows local cultural attitudes towards foreigners are enormously persistent (e.g. Becker and Pascali, 2019; Becker et al., 2016; Voigtlander and Voth, 2012, 2015). We extend this literature by showing that volatile, short-lived bursts in sentiment *within* a given location have substantial effects on people's behavior and that social media may play a role in their propagation. Our findings are particularly timely in light of recent policy debates about whether and how to "regulate" hate speech on social media. Such legislation may come at a high price: since the lines between what constitutes free speech and hate speech can be blurred, regulation can open the door to censorship. Our work does, however, suggest that policymakers ignore online hate speech at their peril. Future research should investigate effective ways to tackle online hate speech. By quantifying the extent of the challenge, our paper takes a first step towards identifying potential harm arising from extended social media usage.

1.5 Appendix: Fanning the Flames of Hate

1.5.1 Appendix: A Short History of the AfD

The AfD was founded by Bernd Lucke, a professor of Economics at the University of Hamburg in 2013. Initially, the AfD positioned itself as an opposition party to the common European currency and the bailouts Greece and Spain received as a result of the financial crisis. Right from the start, however, the party also pandered to the right with a conservative social policy. Representatives of the AfD frequently attracted attention for using nationalist terminology and attacking the "Lügenpresse" (Lying Press), a term popularized by the Nazis. With this political program and rhetoric, the AfD attracted 4.7% of the votes in the 2013 German Federal Election, only narrowly missing the 5% electoral threshold.

Nonetheless, the AfD celebrated several victories in state elections and winning seats in the state parliaments of Hesse, Saxony, Thuringia, Brandenburg, Bremen, and Hamburg. Furthermore, the AfD reached 7.1% of the votes in the 2014 European Parliament election. As the Euro Crisis cooled, the party began to shifts its focus further to the right on topics like traditional family values or the role of Islam in Germany. These more nationalist-conservative political positions, championed by Frauke Petry, attracted a significant share of far-right recruits to the party. In 2015, Petry was elected the main speaker of the party, a major defeat for its founder, Bernd Lucke. As a result of this loss, Lucke resigned from his leadership position and left the party completely, followed by several other key party members.

In the run-up to the 2017 federal election, the AfD leadership included Frauke Petry, Alexander Gauland, Björn Höcke, Jörg Meuthen, and Beatrix von Storch, all of whom hold staunch national conservative opinions. With the beginning of the refugee crisis, the aggressively framed mass immigration as dangerous and declared they were unwilling to accept any refugees into Germany. This messaging was accompanied by increased xenophobia and criticism of Islam.

Under the new leadership and impelled by the refugee crisis, the AfD continued to win elections, securing seats in 14 out of the 16 state parliaments in 2016. In the 2017 federal election, the AfD became the third strongest force in the German Parliament with 12.6% of the votes.

1.5.2 Appendix: Additional Details on the Data

A Short Introduction to Facebook Pages and User Data

On Facebook, celebrities, universities, restaurants, and political groups like the AfD have created their own Facebook pages. The AfD page is the starting point for its followers on Facebook. Any Facebook user who is interested in or supports the AfD can "like" its page. The messages posted on the AfD's page then will show up in that user's Facebook feed. The Facebook feed consists of the individualized news and updates every user receives based on his friendship network and interests. In this way, the AfD is able to reach and rally their followers with political messages and party news.

In addition to receiving information from the AfD, Facebook users can become active on the party's page as well. In general, such interactions fall into three categories. First, people can post their own messages, links, or pictures on the fan page. These posts are visible to everybody but will not automatically appear in other users' Facebook feeds. Second, users can comment on posts and comments by other users or the AfD itself. Those comments appear below the original post and are also visible to the public. Third, each post or comment can be "liked" as a sign of support.

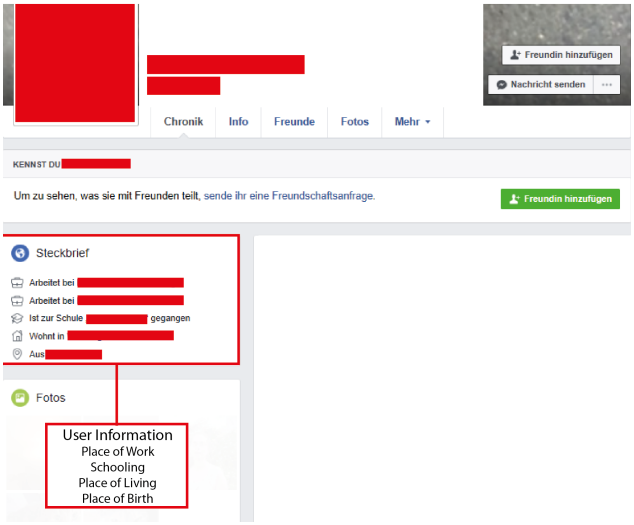
Figure 1.8a shows an example of how these three interaction types show up on the AfD page. The Facebook Graph API allowed us to collect all post, comments, and likes from the AfD's fan page, information we highlight ins Figure 1.8a. Facebook assigns each user a unique ID that makes it possible to attribute posts and comments to individual profiles.

To hand collect user data, one must visit each individual Facebook user profile, from which, depending on the user's privacy settings, one can determine his/her place of residence and place of birth. Figure 1.8b shows an example of a Facebook user profile and where to find the relevant information. If the user decided to hide this information, the box with the user information will be empty.

Figure 1.8: Facebook Examples

(a) Alternative for Germany Facebook Page

(b) Facebook User Profile



Notes: Panel (a) shows a screenshot of the Alternative for Germany's Facebook page. The boxes and labels highlight the parts extracted using the Facebook Graph API. Panel (b) shows an example of a Facebook user profile. The box highlights the publicly available user information extracted from Facebook. The authors removed users' personal information for privacy.

Table 1.8: Translated Example AFD Posts From Facebook

Date	Post	# Likes	# Comments	# Shares
19/05/2017	Side note in the local newspaper: A Turkish man (23) raped a young woman for more than four hours and was cleared of all charges by the judge. Verdicts that were only known in Arabic cultures are now finding their way into Germany. These pro multi cultural diversity judges are raping the German justice system, cultural sensitivity is apparently more important than the rule of law.	18917	302	1
10/05/2017	+++EUR 204.5 million per month for 500,000 asylum seekers paid in unemployment benefits+++ The top politicians of the old parties sold us the wave of migrants coming to Germany as an enrichment of culture and the economy, but the reality looks very different. The former skilled workers are being financed by social security because they cannot get a job because they are uneducated. Deportations are still not enforced and as a result everybody is fed by the state, even those without asylum.	2418	299	1446
27/12/2016	In Berlin, the police has arrested the wanted teenagers who are suspected to have set a homeless person on fire. Out of the seven suspects between the ages of 15 and 21, six are from Syria and one from Lybia. According to a report of the *Süddeusche Zeitung*, all of them arrived in Germany as "refugees" between 2014 and 2016.	7984	1665	5725
15/11/2016	+ We knew it: More "refugees" are now coming via plane + The government currently has 500 migrants per month flown in from Italy and Greece. The Minister of Interior is further reviewing the admission of an additional 13,500 refugees from Turkey. Only a few European countries are complying with the EU directives, Germany - how could it be any different - is one of the first in line.	2153	1066	2584
21/10/2016	+++ Civil war in Garmisch-Partenkirchen? +++ The *Süddeutsche Zeitung* reports that the situation in Garmisch-Partenkirchen seems to be disastrous: "Blacks have taken over the power in the small skiing village in Germany", the Kremlin-financed Russian station Russia Today reports. The French right-wing news portal Atlantic reports similar things about the alleged regime of dark-skinned refugees and the British Daily Mail speaks of riots in the streets , vandalism, and open sexual assaults.	2084	698	1926

Notes: This table reports five example posts from the AfD Facebook page that were posted by the party itself. The post were translated by the authors.

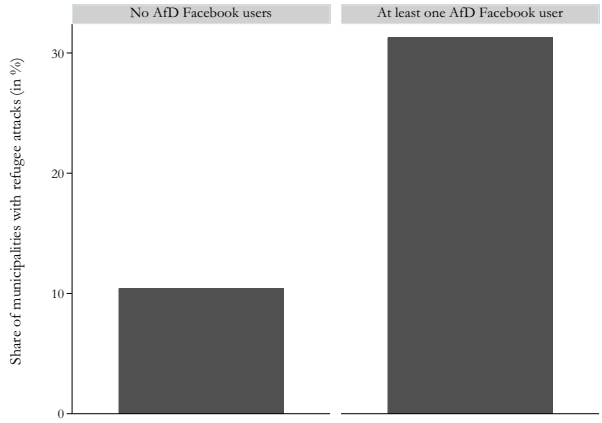
Examples on Anti-Refugee Incidents

Table 1.9: Examples of Anti-Refugee Incidents

Date	Place	Description	Type
03.11.2016	Braunsbedra	A 20-year old Syrian man was riding his bike in the evening and approached a man he assumed needed his help. Suddenly, two additional men appeared and all three started kicking and punching the victim.	Assault
28.12.2016	Langenhagen	An unknown person sprayed graffiti on a refugee home. The graffiti read "Deutsch Nantional" (German-National, misspelled in original), "18" (code for Adolf Hitler) and "88" (code for Heil Hitler).	Property Damage
17.11.2016	Oschersleben	A fire occurred in a villa which had until recently accommodated refugees. After a forensic analysis, the police concluded it to be a case of arson, since the fire started in several places at once using fire accelerant. A detonation occurred when the police arrived. Nobody was injured.	Arson
30.01.2016	Schmölln	450 people participated in a demonstration of the "Thügida" (Pegida in Thuringia). The police charged 4 people with violating gun control laws and the Public Meetings Act.	Demonstration
30.01.2016	Berlin	The police investigated an insult against inhabitants of a refugee home.	Suspected Cases

Notes: This table reports one example for each class of anti-refugee incidents in the data. The descriptions were translated by the authors.

Figure 1.9: Share of Municipalities With Refugee Attacks, by AfD Users



Notes: This figure plots the share of municipalities with at least one refugee attack in our sample by whether we have evidence of at least one AfD Facebook page user in the municipality. We are able to identify one or more AfD users in 3,563 municipalities; for 903 municipalities we find no AfD user.

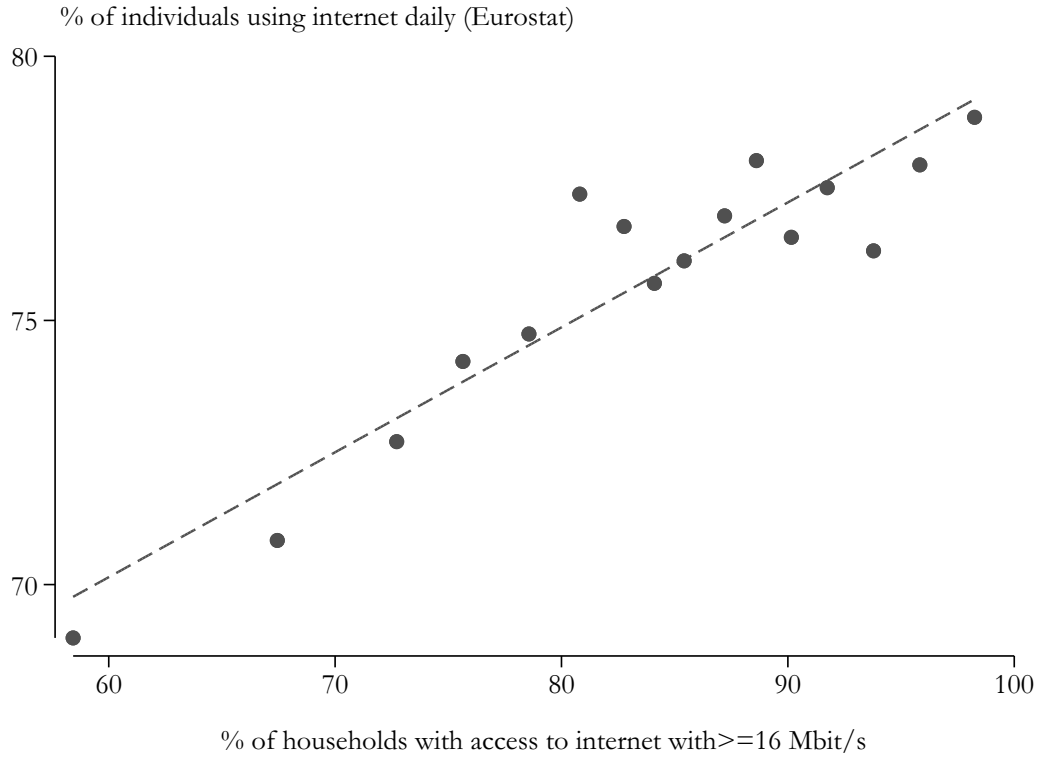
Additional Variable Overview

Table 1.10: Summary Statistics for Additional Controls

	Level	Obs	Mean	SD	Min.	Max.
Additional Media and Internet Controls						
Internet outages/Pop.†	Municipality	495,726	0.007	0.071	0.000	3.614
Registered domains/Pop.	County	495,726	0.141	0.056	0.057	1.390
Mobile broadband speed	County	495,726	12	2	6	24
News paper sales/Pop.	Municipality	491,175	0.092	0.077	0.000	1.644
Additional Right Wing Controls						
Nazi murders/Pop.†	Municipality	495,726	0.002	0.012	0.000	0.282
NPD vote share (2017) (in %)	Election Distr.	495,726	0.491	0.410	0.000	2.006
Ln(1+Deported Jews)	Municipality	495,726	0.606	1.350	0.000	10.930
Ln(1+Stürmer letters)	Municipality	495,726	0.125	0.449	0.000	5.872
Additional Soci-Economic Controls						
Average age	Municipality	479,853	45	2	27	56
Share benefit recipients (in %)	Municipality	495,726	0.424	0.187	0.051	1.207
Share non-Christians	Municipality	479,853	0.325	0.219	0.029	0.941
Manufacturing share (in %)	County	493,062	27	9	2	58
Additional Voting Controls						
CDU vote share (in %)	Municipality	492,618	36	7	20	64
SPD vote share (in %)	Municipality	492,618	19	7	5	47
Left vote share (in %)	Municipality	492,618	8	4	2	26
Green vote share (in %)	Municipality	492,618	7	4	1	25
FDP vote share (in %)	Municipality	492,618	10	3	3	28
Voter turnout (in %)	Election Distr.	495,726	76	3	66	84
Additional Demographic Controls						
Share aged 0-24 (in %)	Municipality	479,853	25	3	14	37
Share aged 25-49 (in %)	Municipality	479,853	33	2	22	45
Share aged 50-74 (in %)	Municipality	479,853	33	3	22	50
Share aged 75 and older (in %)	Municipality	479,853	9	2	4	18

Notes: This table reports summary statistics for the additional control variables in the estimation sample. Variables tagged with a † are scaled by population (in 1,000).

Figure 1.10: Daily Internet Users and Share of Households with Broadband Access



Notes: This figure plots the municipal-level share of households with access to broadband internet ($\geq$ 16 Mbit/s) against the state-level percentage of individuals using the internet daily taken from Eurostat survey data, binned into 16 quantiles. The corresponding correlation coefficient is 0.9245.

Table 1.11: Overview Variables

(a) Part 1/4

Variable	Level	Description	Source
		Refugee Attacks	
Refugee Attacks/Refugees	Muni.-Week	Constructed by dividing the number of anti-refugee incident in a municipality and week by the number of refugees.	Amadeu Antonio Foundation
Arson Attacks/Refugees	Muni.-Week	Same as Refugee Attacks/Refugees but limited to arson attacks as classified by the Amadeu Antonio Foundation.	Amadeu Antonio Foundation
Other Property Attack/Refugees	Muni.-Week	Same as Refugee Attacks/Refugees but limited to attacks leading to miscellaneous property damages as classified by the Amadeu Antonio Foundation.	Amadeu Antonio Foundation
Assaults/Refugees	Muni.-Week	Same as Refugee Attacks/Refugees but limited to assault as classified by the Amadeu Antonio Foundation.	Amadeu Antonio Foundation
Demonstrations/Refugees	Muni.-Week	Same as Refugee Attacks/Refugees but limited to demonstrations as classified by the Amadeu Antonio Foundation.	Amadeu Antonio Foundation
Suspected Cases/Refugees	Muni.-Week	Same as Refugee Attacks/Refugees but limited to suspected attacks still under investigation, as classified by the Amadeu Antonio Foundation.	Amadeu Antonio Foundation
		Social Media Data	
AfD Users/Pop.	Municipality	The number of AfD Users in each municipality divided by population.	Facebook
Refugee Posts	Week	The number of posts on the AfD Facebook page that contain the word 'Flüchtling' (refugee) in a given week.	Facebook
Posts/AfD Users	Municipality	The total number of posts attributed to AfD users of a given municipality divided by the number of AfD Facebook users.	Facebook
Comments/AfD Users	Municipality	Total number of comments that posts by AfD users of a given municipality received divided by the number of AfD Facebook users.	Facebook
Likes/AfD Users	Municipality	The total number of likes that posts by AfD users in a given municipality received divided by the number of AfD Facebook users.	Facebook

(b) Part 2/4

Variable	Level	Description	Source
		Auxiliary Variables	
$I_{Internet\ Outage}$	Muni.-Week	Dummy variable that is equal to 1 for municipality-week observations that are in the top quartile of the reported internet outages per capita ratio, and 0 otherwise. The number of user-reported outages comes from Heise.de. We exclude outages that are shorter than 24 hours.	Heise.de
$I_{Facebook\ Outage}$	Week	Dummy variable that is equal to 1 for the weeks with major Facebook outages as described in Table 1.13, and 0 otherwise.	Various news sources
		Baseline Controls	
Population	Municipality	The population of each municipality in 2015 from the shape file of the BKG. The population numbers in the shape file are equivalent to the 2015 data from the German Statistical Office (Destatis).	BKG/Destatis
GDP/Worker	County	GDP per working population at the county-level.	Destatis
Population density	Municipality	Population density, defined as population over municipality size (in km^2).	Destatis
AfD vote share (2017)	Municipality	The share of votes cast for the AfD in the 2017 German Federal Parliament Election.	Destatis
Share high school	Municipality	The share of people whose highest educational attainment is at least "Abitur", the German high-school certificate.	Destatis
Share Broadband access	County	The share of the population that have access to at least 16 Mbit/s internet connection speed.	© BMVI, TÜV Rheinland
Share immigrants	Municipality	The share of the population that are immigrants.	Destatis
Asylum Seekers/Pop.	County	The asylum seekers per capita.	Destatis
		Raw Data	
Refugee attacks	Muni.-Week	The number of anti-refugee incident in a municipality and week.	Amadeu Antonio Foundation
Population (2015)	Municipality	The population for each municipality.	BKG
Refugees (2015)	County	The number of asylum seekers in each county.	Destatis
AfD Users	Municipality	The number of users of the AfD Facebook page we could locate based on their reported place of residence.	Facebook

(c) Part 3/4

Variable	Level	Description	Source
Additional Media and Internet Controls			
Internet outages/Pop.	Municipality	The total number of large internet outages per capita (as defined above) that occurred in a municipality over the sample period	Heise.de
Registered domains/Pop.	County	The number of registered .de domains in a given county, divided by the county population.	Destatis
Mobile Broadband Speed	County	The average mobile download speed in Mbits/s.	BMVI
Newspaper sales/Pop.	Municipality	The number of newspaper copies sold in a given municipality, divided by the population. The data do not contain information for municipalities smaller than 3000 inhabitants, which we impute using the population, population density, AfD vote share, and county fixed effects (results are almost equivalent without imputation).	ZMG
Additional Right-Wing Controls			
Nazi murders (1990-2016)	Municipality	The number of murders classified as having a neo-Nazi motive in a municipality between 1990 and 2016, scaled by population.	Mut gegen rechte Gewalt
NPD vote share (2017)	Election Distr.	The share of votes cast for the extremist right-wing NPD (National Democratic Party of Germany) in the 2017 German Federal Parliament Election.	Bundeswahlleiter
Log(1+Deported Jews)	Municipality	The natural logarithm of the number of Jews who were deported during Nazi times. To analyze cross-sectional correlates, we scale the number of deported Jews by the number of Jews in a municipality in 1933 (see text for details).	Voigtlander and Voth (2012)
Log(1+Stürmer letters)	Municipality	The natural logarithm of the number of letters written to "Der Stürmer", the anti-Semitic newspaper published by Nazi politician Julius Streicher. To analyze cross-sectional correlates, we scale the number of letters by the population in 1933 (see text for details).	Voigtlander and Voth (2012)
Additional Socio-Economic Controls			
Average age	Municipality	The average age in each municipality.	Destatis
Benefit recipients/Pop.	Municipality	The number of social benefit recipients in a given municipality divided by the population.	Destatis
Non-christians/Pop.	Municipality	The number of non-Christians in a given municipality divided by population.	Destatis
Manufacturing share (%)	County	The share of manufacturing employees in a given county.	Destatis

(d) Part 4/4

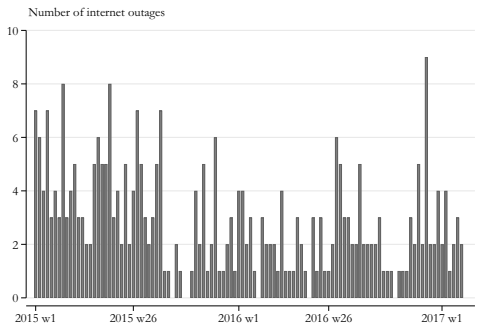
Variable	Level	Description	Source
Additional Voting Controls (2013 & 2017 Election)			
CDU vote share	Municipality	The share of votes cast for the CDU in the 2017 German Federal Parliament Election.	Destatis
SPD vote share	Municipality	The share of votes cast for the SPD in the 2017 German Federal Parliament Election.	Destatis
Left vote share	Municipality	The share of votes cast for "Die Linke" (The Left) in the 2017 German Federal Parliament Election.	Destatis
Green vote share	Municipality	The share of votes cast for the party "B90/Die Grünen" (Green Party) in the 2017 German Federal Parliament Election.	Destatis
FDP vote share	Municipality	The share of votes cast for the FDP in the 2017 German Federal Parliament Election.	Destatis
Pirate vote share	Municipality	The share of votes cast for the Pirate party in the 2017 German Federal Parliament Election.	Destatis
Voter turnout	Election Distr.	The voter turnout in the 2017 German Federal Parliament Election.	Bundeswahlleiter
Additional Demographic Controls			
Share aged 0-24	Municipality	The number of people aged 0-24, divided by population.	Destatis
Share aged 25-49	Municipality	The number of people aged 25-49, divided by population.	Destatis
Share aged 50-74	Municipality	The number of people aged 50-74, divided by population.	Destatis
Share aged above 75	Municipality	The number of people aged 75 and up, divided by population.	Destatis

1.5.3 Appendix: Additional Details and Results on Internet and Facebook Outages

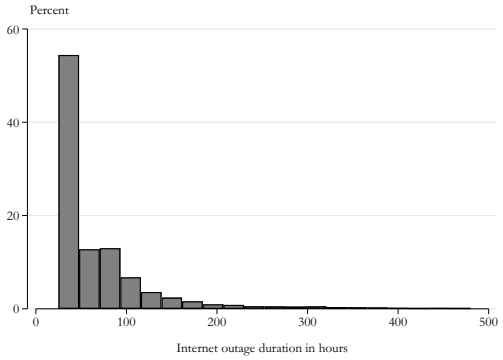
Figure 1.11: Spatial and Temporal Distribution of Internet Outages

(a) Map of Internet Outages

(b) Internet Outages Over Time



(c) Duration of Internet Outages



Notes: The map in Panel (a) plots the geographic distribution of internet outages per million inhabitants for the German municipalities in the data. Panel (b) plots the distribution of total internet outages per week. Panel (c) plots the distribution of the duration of the individual user reports from *Heise.de* used in the regressions, trimmed at three weeks. See Section 1.2 for more details.

Table 1.12: Validation of Internet Outage Data

(a) Part 1/2

Date	Provider	Region	Description	# Outages	Source
12/06/2015	Kabel D. and Unitym.	Germany	The IT website "Netzwelt" reported a large internet outage on June, 12th 2015. Users of the providers Kabel Deutschland and Unitymedia were especially affected. According to a spokesperson for Kabel Deutschland, the problem was caused by a disruption at the internet hub in Frankfurt.	61	Link
18/06/2015	Unitymedia	Karlsruhe	On July, 18th 2015, the news site "KA News" reported a disruption at internet provider Kabel BW, a subsidiary of Unitymedia. Kabel BW confirmed the problem and explained that their technicians were currently working to fix the problem. The outage affected the area of Karlsruhe.	36	Link
24/06/2015	Unitymedia	Cities in NRW	The "Rheinische Post" reported on June 24th, 2015 that many users of the provider Unitymedia encountered disrupted internet connections beginning on Wednesday, June 23th. Most of the reports came from the cities of Düsseldorf, Mönchengladbach, Neuss, and Münster. Unitymedia did not provide an official statement.	15	Link
05/07/2015	O2 and 1u1	Berlin	The IT website "Golem.de" reported on July, 5th 2015 that users of DSL provider 02 and 1und1 reported disruptions of their internet and phone connections. The problems had started on the June 27th and were largely fixed by the evening of the 5th. Neither provider explained what had caused the problems.	27	Link
08/07/2015	Versatel	Münster	The "Halterner Zeitung" reported on July 8th, 2015 that households in the city of Haltern were cutoff from the internet. The outage was caused by a damaged fiber optic cable. The same cable was also used by internet provider Unitymedia. As a result, Unitymedia users in the Münster area were also affected by the problem.	29	Link
20/08/2015	Unitymedia	NRW and Hessen	The "Gießener Allgemeine" reported on August, 20th 2015 that many users of internet provider Unitymedia encountered disrupted internet connections beginning August 19th. The internet outage affected the entire state of Nordrhein-Westfalen as well as parts of Hessen. At the time of the report, Unitymedia was still investigating the cause of the outage.	81	Link

(b) Part 2/2

Date	Provider	Region	Description	# Outages	Source
04/12/2015	Telekom	Major cities	"ZDnet.de", a website specialized in IT and electronics, reported on December 4th, 2015 that users of the internet provider German Telekom were encountering disrupted internet connections beginning in the early morning of the same day. Most of the reports came from the major cities Berlin, Hamburg, Munich, and Frankfurt. According to the German Telekom, the problem was caused by a breakdown of a RADIUS server that is responsible for authenticating internet access.	19	Link
30/06/2016	Vodafon and Kabel D.	Germany	"Heise.de", the website from which we obtained the outage data, reported on June 30th, 2016 an outage of the internet provider Kabel Deutschland that affected the entirety of Germany. The outage was caused by a problem at a computer cluster. At the time of the report the outage was still ongoing.	122	Link
21/07/2016	T-Online	Germany	The IT website "Golem.de" reported on July ,21th 2016 on internet problems with the provider T-Online. The outage affected not only private households but also business customers of T-Online. A representative of the provider confirmed the problems but did not name any specific cause. At the time of the report technicians were still working to fix the problem.	41	Link
24/11/2016	O2	Germany	On November 24th, 2016 the website "Chip.de" reported a Germany-wide outage of the internet provider O2. The problems were concentrated in metropolitan areas. At the time of report O2 was still investigating the cause of the problem, which was likely an issue with the company's VoIP system.	31	Link
27/11/2016	Telekom	Ruhr area	The "Spiegel" reported on November 27th, 2016 that users of the internet provider German Telekom were cut off from the internet. The outage mainly affected the Ruhr area, but internet problems were also reported in Frankfurt, Hannover, and Braunschweig. At the time of the report Telekom was still working to correct the problem.	19	Link

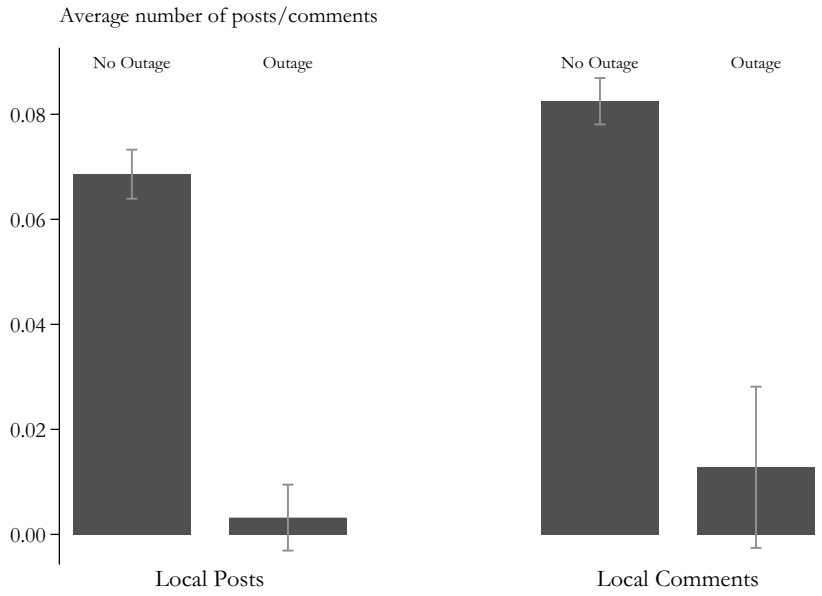
Notes: This table reports several examples of internet outages that were reported in German newspapers or on well-known specialized websites. Each entry lists the date of the outage as well as the affected provider and region. The table further features a short description of the outage and a link to the original news source. The column "# Outage" refers to the number of outages of the affected provider reported by users on Heise.de, which serve as the basis of our internet outage measures; note that this number reflects the number of user reports, *not* the actual number of affected users. The web pages were last accessed in February 2018.

Table 1.13: Validation of Facebook Outage Data

Peak Date	Description	Source
26/01/2015	The Facebook page was unavailable globally due to a server error. According to the official statement, the error "occurred after we introduced a change that affected our configuration systems." Initially, the outage had been attributed to an attack by infamous hacker group "Lizard Squad". The outage affected millions of users worldwide, including users of Facebook messenger, Instagram, and the dating app Tinder (which uses Facebook data).	Link
08/04/2015	Facebook users complaint that the site is not loading properly. The outage was particularly concentrated in Western Germany, the Netherlands, and the United Kingdom.	Link
15/07/2015	Facebook suffered a worldwide outage, showing users a simple "Service Unavailable" message. The outage affected all services including the popular Facebook messenger. Although the initial issue was resolved relatively quickly, the problems persisted for many users.	Link
29/09/2015	Users experienced extremely slow or no access to Facebook after a previous disruption on September 24. User reports and news coverage indicate that Germany was particularly badly hit. In a statement to CNBC, Facebook acknowledged the outage and explained that "configuration problems" were at the root of it.	Link, Link 2
14/03/2016	Users in Western Europe - particularly Germany, Austria, Poland, the Netherlands, Belgium, and the United Kingdom - were barred from logging into or commenting on Facebook. The Facebook app was particularly affected.	Link
16/06/2016	Facebook had an outage concentrated in Western Europe. Users were unable to log in, post, use the messenger, or could not access pages (including that of the AfD).	Link
14/09/2016	Worldwide Facebook outage, affecting almost the entire European continent and the eastern United States. Users were unable to log in, post, or read content.	Link
13/01/2017	Users in Western Europe and the eastern United States experienced widespread issues in accessing Facebook, particularly from computer devices.	Link, Link 2

Notes: This table lists the dates of the major Facebook outages that occurred during our sample period. The links lead to the news articles used to identify the disruptions.

Figure 1.12: Do Local Internet Outages Reduce Local Facebook Activity?



Notes: This figure plots the arithmetic mean and 95% confidence intervals of local Facebook activity measures based on linking users' locations to their posts and comments. The bars marked "Outage" are municipality-week observations in which a local internet outage occurs. The average values are below one, since we do not observe a post or a comment from each municipality in every week. For example, a mean value of around 0.08 for local comments during weeks without outages implies that on average we observe 1 comment for every 12.5 municipality-weeks pairs (out of 480,963) in our data.

Table 1.14: Time Series Evidence — Outages and Aggregate Facebook Activity

	Post outcomes				Outage correlation
	(1) Total posts	(2) Refugee posts	(3) Total posts	(4) Refugee posts	(5) Facebook outage (t)
Facebook outage (t+1),	46.159	-9.413			
	(78.768)	(17.718)			
Facebook outage	-107.372**	-19.880**			
	(50.326)	(7.917)			
Facebook outage (t-1),	-26.517	-15.459*			
	(78.355)	(7.895)			
Internet outage			8.178	0.739	-0.007
			(12.093)	(1.728)	(0.017)
Observations	108	108	109	109	111
R-squared	0.368	0.830	0.344	0.812	0.002
Week-of-year FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: This table presents weekly time series regressions of different metrics of Facebook activity on a dummy for Facebook outages in a given week (columns 1-2). Columns 3-5 show that the sum of local internet outages is uncorrelated with Germany-wide Facebook posts and the likelihood of a severe Facebook outage. Newey-West standard errors are reported in parentheses. ** and * indicate statistical significance at the 0.05 and 0.1 level, respectively.

Table 1.15: Robustness — Ruling Out Alternative Channels

	(1) Official Reports	(2) Leave One Out Estimator	(3) Lagged Posts	(4) Google Sentiment Measure
AfD users/Pop. × Refugee posts	0.009*	0.057***	0.011	0.103***
	(0.005)	(0.021)	(0.008)	(0.032)
AfD users/Pop. × Posts × Outage	-0.137***	-0.372***	-0.164***	-0.571***
	(0.042)	(0.115)	(0.062)	(0.219)
Observations	474,303	474,303	470,030	474,303
R-squared	0.045	0.084	0.084	0.084
Municipalities	4273	4273	4273	4273
Municipality FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
All controls [30] × Posts	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). Column 1 only uses anti-refugee incidents based on official reports (police or parliament), which are unlikely to be subject to time-varying reporting bias. In column 2 we construct a leave one out measure of *Refugee posts*. Internet outages are defined as municipality-weeks that are in the top quartile of the ratio of reported internet outages to population. Columns 1-3 include all controls as in column 7 of table 1.2, interacted with *Refugee posts* and all additional interactions of the outage dummy (unreported). Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Table 1.16: Outage Results with Alternative Standard Errors

Panel A: Internet Outages

	(1) Municipality	(2) Municipality and week	(3) County	(4) County and week	(5) Driscoll-Kraay BW(4)	(6) Driscoll-Kraay BW(10)	(7) Conley 20km	(8) Conley 50km	(9) Conley 100km
AfD users/Pop. × Refugee posts	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***
	(0.009)	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)	(0.005)	(0.005)	(0.006)
AfD users/Pop. × Posts × Outage	-0.181***	-0.181***	-0.181***	-0.181***	-0.181***	-0.181***	-0.181***	-0.181***	-0.181***
	(0.058)	(0.059)	(0.056)	(0.057)	(0.064)	(0.060)	(0.063)	(0.063)	(0.063)

Panel B: Facebook Outages

	(1) Municipality	(2) Municipality and week	(3) County	(4) County and week	(5) Driscoll-Kraay BW(4)	(6) Driscoll-Kraay BW(10)	(7) Conley 20km	(8) Conley 50km	(9) Conley 100km
AfD users/Pop. × Refugee posts	0.027***	0.027***	0.027***	0.027***	0.027***	0.027***	0.027***	0.027***	0.027***
	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)	(0.009)	(0.006)	(0.006)	(0.006)
AfD users/Pop. × Posts × Outage	-0.040*	-0.040*	-0.040*	-0.040*	-0.040*	-0.040*	-0.040**	-0.040**	-0.040**
	(0.021)	(0.021)	(0.023)	(0.023)	(0.024)	(0.023)	(0.020)	(0.020)	(0.017)
Observations	479,964	479,964	479,964	479,964	479,964	479,964	479,964	479,964	479,964
R-squared	0.082	0.082	0.082	0.082	0.082	0.082	0.002	0.002	0.002
Nr. of Clusters	4324	4324/111	394	394/111	4324/111	4324/111			
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.2). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee("Flüchtling"). Internet outages are defined as municipality-weeks in the top quartile of the internet outage reports to population ratio. Robust standard errors are constructed as defined in the top row. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Table 1.17: Outage Results with Alternative Functional Forms

Panel A: Internet Outages

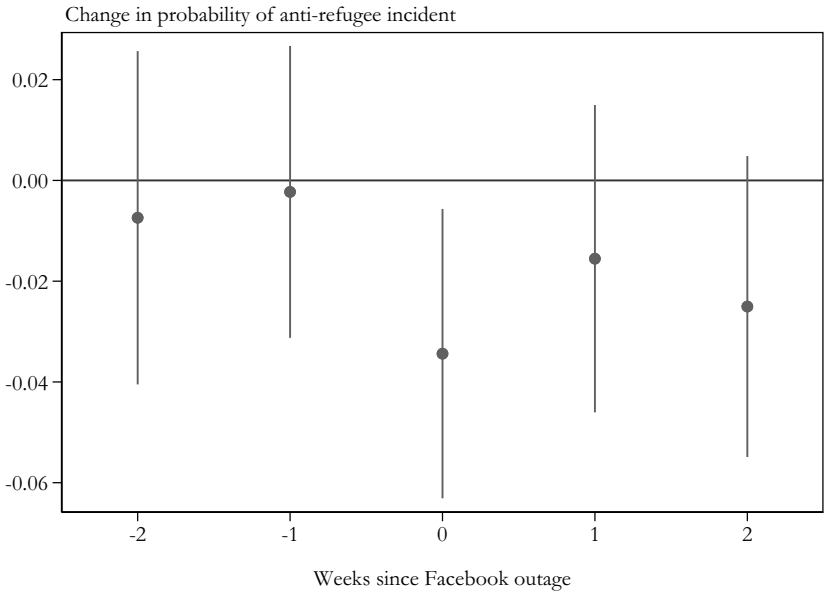
	Refugee Attack Dummy		Refugee Attacks		Log(1 + Refugee Attacks)		Refugee Attacks/Asylum Seeker	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AfD users/Pop. × Refugee posts		0.016**		0.017*		0.012*		0.018**
		(0.008)		(0.010)		(0.006)		(0.008)
AfD users/Pop. × Posts × Outage		-0.184***		-0.187***		-0.130***		-0.192**
		(0.058)		(0.065)		(0.042)		(0.082)
Outage	-0.004***	-0.002	-0.004***	-0.003	-0.003***	-0.002	-0.002***	-0.002
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
Observations	474,303	474,303	474,303	474,303	474,303	474,303	474,303	474,303
R-squared	0.084	0.084	0.157	0.157	0.120	0.120	0.046	0.046
Municipalities	4273	4273	4273	4273	4273	4273	4273	4273
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
All controls [30] × Posts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Facebook Outages

	Refugee Attack Dummy		Refugee Attacks		Log(1 + Refugee Attacks)		Refugee Attacks/Asylum Seeker	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AfD users/Pop. × Refugee posts		0.021**		0.022**		0.015**		0.022**
		(0.009)		(0.010)		(0.007)		(0.009)
AfD users/Pop. × Posts × Outage		-0.046**		-0.062		-0.036*		-0.039**
		(0.022)		(0.040)		(0.019)		(0.019)
AfD users/Pop. × Outage		1.367		2.700		1.234		0.242
		(1.862)		(3.218)		(1.527)		(1.426)
Outage	-0.001***		-0.002***		-0.001***		-0.002***	
	(0.000)		(0.000)		(0.000)		(0.000)	
Observations	474,303	474,303	474,303	474,303	474,303	474,303	474,303	474,303
R-squared	0.081	0.084	0.155	0.157	0.117	0.120	0.045	0.046
Municipalities	4273	4273	4273	4273	4273	4273	4273	4273
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week FE		Yes		Yes		Yes		Yes
All controls [30] × Posts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage, anti-refugee sentiment, and outages as in Equation (1.2). The dependent variable is the measure of refugee attacks listed in the top row. *Refugee Attacks/Asylum Seeker* is the number of anti-refugee incidents per 1000 asylum seekers. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee("Flüchtling"). Note that the interactions of internet outages with refugee posts and AfD Facebook usage are included but unreported to save space. Robust standard errors are constructed as defined in the top row. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Figure 1.13: Facebook Outage Event Study



Notes: This figure plots estimates the estimates for λ from an event study regression of Equation (1.2) which includes 2 leads and lags of the outage interactions. 95% confidence intervals are based on standard errors clustered by municipality.

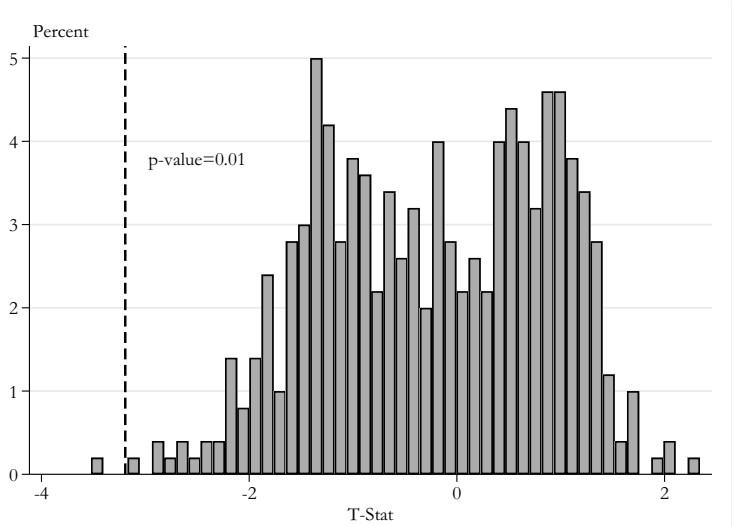
Table 1.18: Robustness — Alternative Definitions of Internet Outages

	Change outage definition		**Include shorter outages**		**No extended outages**	
	(1) (Baseline) 75th percentile	(2) 90th percentile	(3) 75th percentile	(4) 90th percentile	(5) 75th percentile	(6) 90th percentile
AfD users/Pop. × Refugee posts	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
AfD users/Pop. × Posts × Outage	-0.181***	-0.103**	-0.145***	-0.116***	-0.161***	-0.097*
	(0.058)	(0.052)	(0.043)	(0.040)	(0.058)	(0.055)
Observations	479,964	479,964	479,964	479,964	479,964	479,964
R-squared	0.082	0.082	0.082	0.082	0.082	0.082
Municipalities	4324	4324	4324	4324	4324	4324
Number of outages	308	122	579	231	246	98
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.2). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* and textitRefugee posts is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). Each column includes all additional interactions of the outage dummy (unreported). In columns 1-2, we use the baseline dummy explained above, i.e. outages in the top quartile. In columns 3-4, we include outages shorter than 24 hours (as discussed in Section 1.2 we exclude this for our baseline measures) and define a new dummy for outages in the top quartile. Columns 5-6 further do not extended outages beyond a single week. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

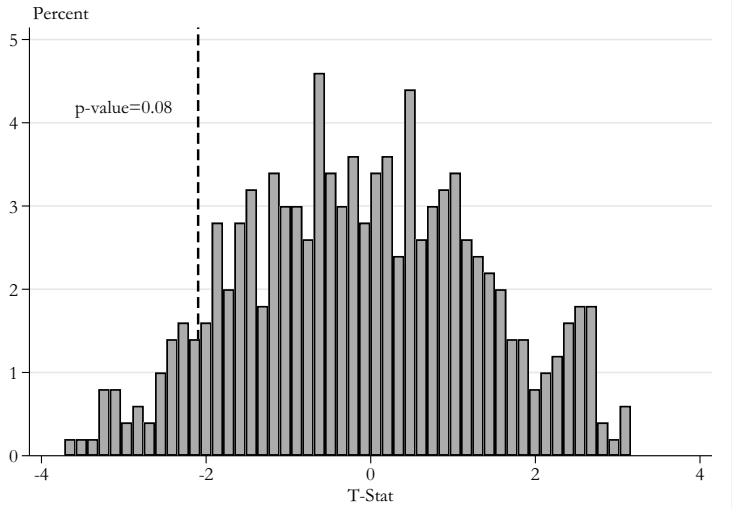
Figure 1.14: Randomization Test for Outage Results

(a) Internet Outages



Notes: This figure shows the results of the randomization test, in which we randomly assign placebo internet outages to 313 municipality-week pairs. We repeat this process 500 times and save the t-stat of the triple interaction term of interest. The vertical line marks the t-stat of the actual estimate.

(b) Facebook Outages



Notes: This figure shows the results of the randomization test, in which we randomly assign placebo Facebook outages to eight weeks in our data. We repeat this process 500 times and save the t-stat of the triple interaction term of interest. The vertical line marks the t-stat of the actual estimate.

1.5.4 Appendix: Additional Results

Table 1.19: Violent vs. Non-Violent Incidents

	(1)	(2)	(3)	(4)
	Arson	Assault	Property Damage	Protest
AfD users/Pop. × Refugee posts	0.002	0.007**	0.012*	0.005*
	(0.002)	(0.004)	(0.006)	(0.003)
Observations	479,964	479,964	479,964	479,964
R-squared	0.016	0.053	0.060	0.060
Municipalities	4324	4324	4324	4324
Municipality FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack; see 1.11 for definition of attack types. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). All control variables are interacted with the *Refugee posts* measure; see text for a description of the controls. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Table 1.20: Other Facebook Posts and Anti-Refugee Hate Crimes

	(1) Refugee (Baseline)	(2) Muslim posts	(3) Islam posts	(4) EU posts
AfD users/Pop. × FB posts	4.027***	0.765	-0.043	0.387
	(1.009)	(0.502)	(0.456)	(0.385)
Observations	495,726	495,726	495,726	495,726
R-squared	0.078	0.078	0.078	0.078
Municipalities	4466	4466	4466	4466
Municipality FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *FB posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"), with the baseline being *refugee* ("Flüchtling"). Standardized coefficients are reported in square brackets, based on variable transformations with a mean of 0 and a standard deviation of 1. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Table 1.21: Social Media Reach and Hate Crime Propagation

	(1) Number of posts	(2) Received comments	(3) Received Likes
AfD users/Pop. × Refugee posts	0.049***	0.048***	0.049***
	(0.017)	(0.017)	(0.017)
AfD users/Pop. × Refugee posts × Reach	0.003***	0.002**	0.001**
	(0.001)	(0.001)	(0.000)
Observations	381,174	381,174	381,174
R-squared	0.086	0.086	0.086
Municipalities	3434	3434	3434
Municipality FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes
Corr(Reach,Population)	0.010	0.011	0.009
Corr(Reach,AfD users/Pop.)	0.017	0.006	0.020

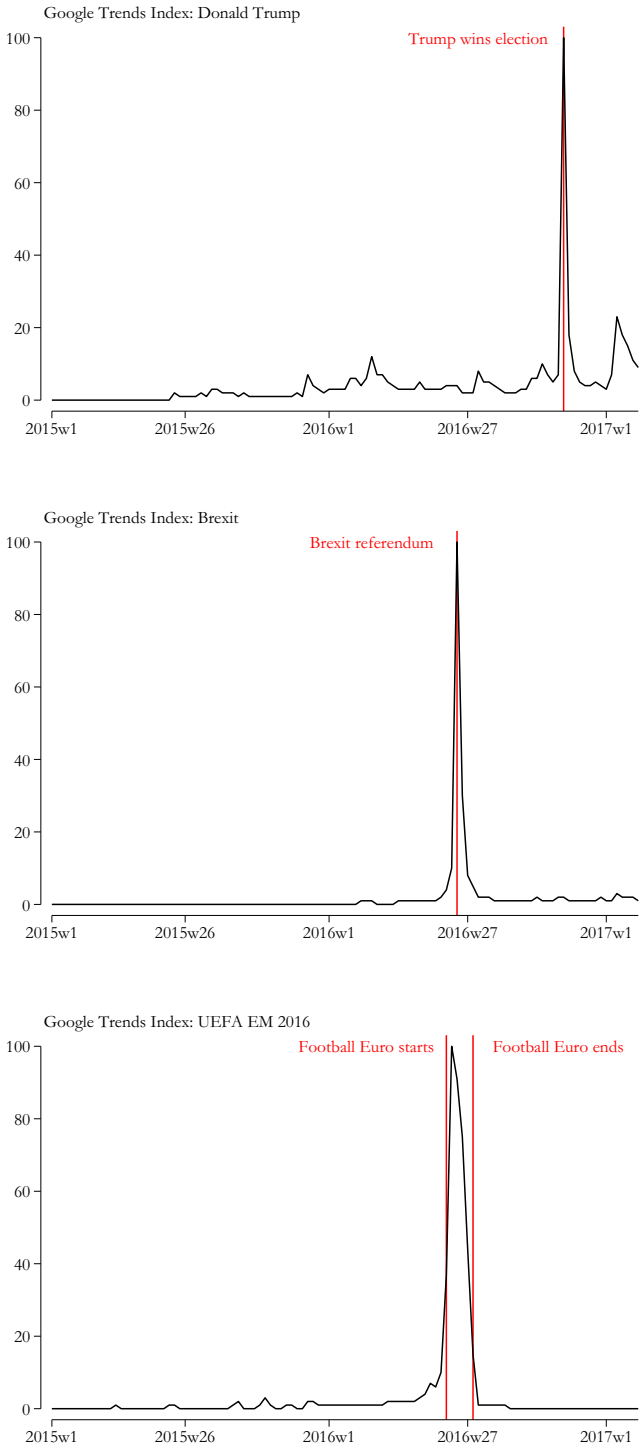
Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). The reach variables in the top row refer to the number of local posts on the AfD wall, as well as comments and likes for AfD posts, all scaled by the number of AfD users (municipalities with zero users are dropped). See text for an explanation of the control variables. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Table 1.22: Time Series Evidence – Distractions and Aggregate Facebook Activity

	(1)	(2)	(3)	(4)
	Brexit	Trump	UEFA EM 2016	Horse race
Brexit	-0.023***			-0.041***
	(0.007)			(0.009)
Trump		-0.020***		-0.018**
		(0.007)		(0.008)
UEFA EM 2016			-0.048***	0.058
			(0.013)	(0.043)
Observations	110	110	110	110
R-squared	0.120	0.119	0.098	0.169
Month FE	Yes	Yes	Yes	Yes

Notes: This table presents weekly time series regressions of the share of refugee posts on the AfD page on Google indeces tracking interest in the topics Brexit, Trump, and the European Football Championship. All regressions include week-of-year and month fixed effects. Newey-West standard errors are reported in parentheses. *** and ** indicate statistical significance at the 0.01 and 0.05 level, respectively.

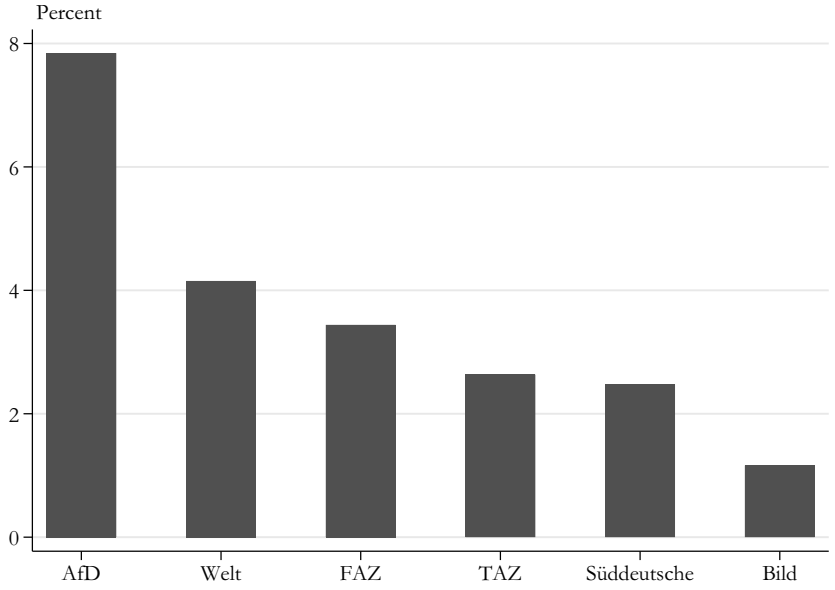
Figure 1.15: Google Trends Data — Brexit, Trump, and Football



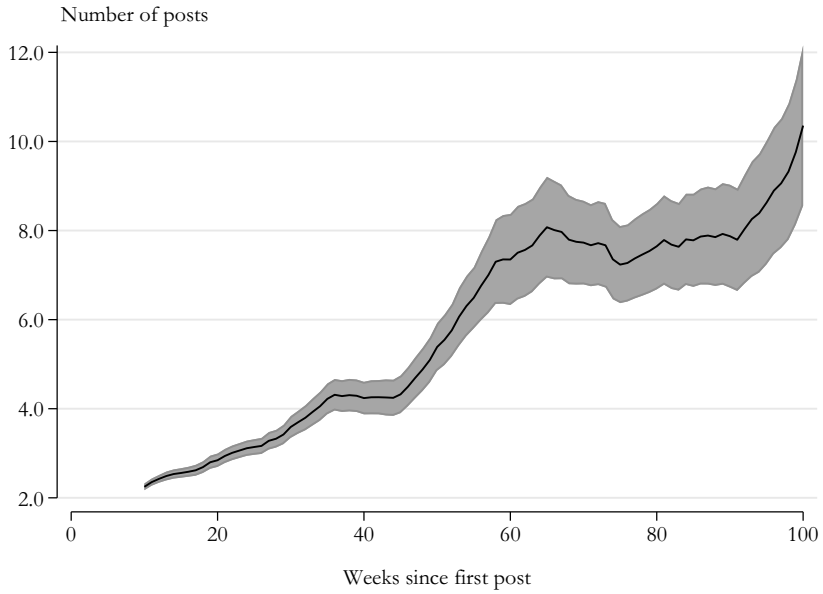
Notes: These figures plot the Google Trends search index (ranging from 0 to 100) for the terms "Trump", "Brexit", and "UEFA EM 2016" for the sample period.

Figure 1.16: Differences of Social and Traditional Media — Additional Results

(a) Share of Posts About Refugees on Facebook (in %), by Page



(b) Individual Posting Behavior, by Length of Exposure



Notes: Figure (a) plots the share of posts on the "wall" of the AfD Facebook page and five major German news outlets. Panel (b) plots the 10-week moving average of the number of posts per person as a function of a user's time spent on the AfD Facebook page, proxied by the time since the first post. The shaded area indicates 95% confidence intervals.

Table 1.23: Relative Word Frequencies on the AfD Facebook Page

Rank	Word	Translation	Relativ prob.
Panel A: Demokra (democratic)			
1	linksliberaldemokratisch	left-liberal	260
2	Demokratiegedanken	idea of democracy	130
3	postdemokratisches	post-democratic	130
4	scheindemokratische	sham democratic	87
5	pseudodemokratischen	pseudo-democratic	65
Panel B: Renter/Kinder/Frauen/Obdachlose (vulnerable groups)			
1	Kinderbande	gang of children	520
2	Kinderbanden	gangs of children	260
3	Burkafrauen	burka women	260
4	Armutsrentnern	poverty-pensioners	260
5	Kindersex	pedophilia	260
Panel C: Elite (elite)			
1	Elitegruppe	elite group	260
2	Leistungseliten	accomplished elites	130
3	Geldeliten	rich elites	87
4	Staatselite	state elites	87
5	Politelite	political elites	74
Panel D: Fremd (foreign)			
1	Fremdkulturen	foreign cultures	676
2	Fremdvölker	foreign people	260
3	Fremdverwendung	foreign use	260
4	fremdgesteuerten	foreign-controlled	130
5	zweckentfremdeter	misused	130
Panel E: Kultur (culture)			
1	Fremdkulturen	foreign cultures	676
2	Kochkultur	cooking culture	520
3	Kulturgewohnheiten	cultural habits	260
4	Rückkehrkultur	return culture	260
5	Clankulturen	clan culture	260

Notes: This table plots the relative probability of words mentioned on the AfD Facebook page compared to reports by major German news outlets on Nexis. We report the results by groups of word stems identified as likely to reflecting right-wing hate speech on social media by previous work in Dinar et al. (2016).

Figure 1.17: Word Cloud — Predictors of AfD Facebook Content

Notes: This figure plots the 100 words with the highest predictive power for AfD Facebook content relative to traditional news media based on an L1 regularized logistic regression model. The size of the words corresponds to the magnitude of the coefficients. The input to the model are the tf-idf weighted words from the AfD's Facebook posts and comments and full text articles from Nexis that contain the word refugee. We exclude words that are shorter than four letters and words that appear in less than five documents. The optimal regularization strength is chosen using 10-fold cross-validation. The selected model achieves an out-of-sample F1 score of above 0.99. The words were translated from German by the authors.

Table 1.24: Mechanism — Local Spillovers

	(1)	(2)	(3)
AfD users/Pop. × Refugee posts	0.024***	0.022***	0.016*
	(0.009)	(0.008)	(0.008)
Attack in neighboring municipality	0.004***	-0.000	0.004**
	(0.001)	(0.002)	(0.002)
Attack in neighboring municipality × Posts		0.000	-0.004**
		(0.001)	(0.002)
Attack in neighboring municipality × AfD users/Pop.		13.765***	1.610
		(4.782)	(4.914)
Attack in neighboring municipality × AfD users/Pop. × Posts			0.121**
			(0.052)
Observations	479,964	479,964	479,964
R-squared	0.082	0.082	0.083
Municipalities	4324	4324	4324
Municipality FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on the AfD Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). *Attack in neighboring municipality* is a dummy equal to 1 if a neighboring town experiences a refugee attack in the same week. The coefficient for "Attack in neighboring municipality × Posts" is multiplied by 100 for readability. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

1.5.5 Appendix: Robustness Checks for Specification

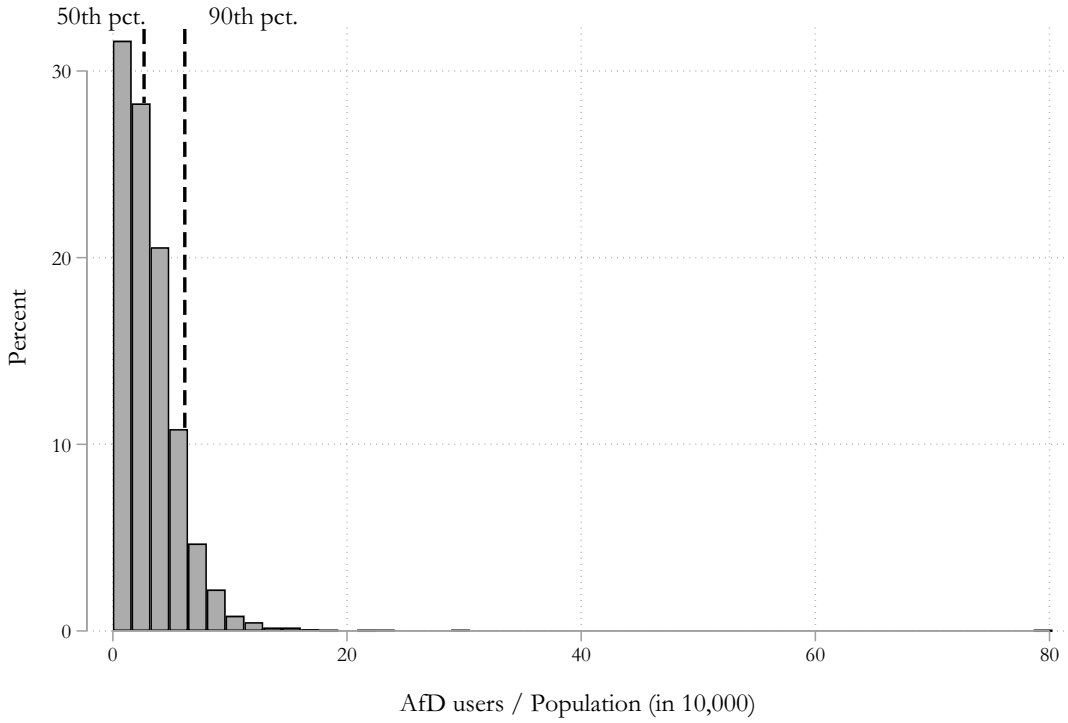
Table 1.25: Further Robustness Checks

	(1) Lagged Dependent Variable	(2) Population Weighting	(3) Pre-2015 Users	(4) Refugee Comments	(5) Refugee Post Likes	(6) Refugeee Post Share
AfD users/Pop. × Refugee posts	0.023***	0.029***	0.034**			
	(0.009)	(0.011)	(0.015)			
AfD users/Pop. × Refugee sentiment				0.076***	0.015**	0.437***
				(0.025)	(0.006)	(0.140)
Observations	475,640	479,964	479,964	479,964	479,964	479,964
R-squared	0.085	0.097	0.082	0.082	0.082	0.082
Municipalities	4324	4324	4324	4324	4324	4324
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls [8] × Posts	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). *AfD users/Pop.* is the ratio of people with any activity on AfD's Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). Column 1 includes a lagged dependent variable. Column 2 uses weighted least squares (WLS) based on each municipality's population. In column 3, we replace the number of AfD users calculated over the whole sample with the number of users before the sample start (that is, pre-2015). Columns 4 and 5 present results based on the comments and likes (in 100s) of posts on the AfD Facebook page containing the word refugee, rather than the number of posts on the Facebook wall. Column 7 uses the share of posts (in %) containing the word refugee in all posts we observe in a given week. All control variables are interacted with the $Refugee\ posts$ measure; see text for a description of the controls. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Accounting for the Skewed Distribution of AfD Users

Figure 1.18: Distribution of AfD Users / Population



Notes: This figure plots the distribution of the ratio of AfD users in a municipality to population. The vertical lines indicate the 50th and 90th percentile of the distribution, respectively, which we make use of in Table 1.26.

Table 1.26: Accounting for the Skewed Distribution of AfD Users

	(1) Drop municipalities 0 users	(2) Only above median	(3) Only below median	(4) 10-90th percentile	(5) User quartiles
AfD users/Pop. × Refugee posts	0.053***	0.035*	0.100***	0.076***	
	(0.019)	(0.020)	(0.026)	(0.024)	
AfD users/Pop. (Q2) × Refugee posts					0.012*
					(0.006)
AfD users/Pop. (Q3) × Refugee posts					0.018***
					(0.007)
AfD users/Pop. (Q4) × Refugee posts					0.057***
					(0.011)
Observations	395,493	247,863	247,863	345,876	395,493
R-squared	0.082	0.097	0.026	0.042	0.082
Municipalities	3563	2233	2233	3116	3563
Municipality FE	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). The dependent variable is a dummy for the incidence of a refugee attack. *AfD users/Pop.* is the ratio of people with any activity on AfD's Facebook page to population. *Refugee posts* is the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). In column 5, the excluded category is the first quartile of *AfD users/Pop.*; zero-user municipalities are excluded. The coefficients in column 5 are multiplied by 1000 for readability. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Addressing Many Zeros in the Dependent Variable

A potential concern with the results in the paper is the sparsity of the dependent variable. Even though anti-refugee incidents were relatively frequent in our sample in absolute numbers, they are rare events in the full sample of municipalities and weeks. In this section, we discuss potential issues and provide evidence for the robustness of our estimates.

The first concern is that the standard errors could be biased, which might make it more likely to find statistically significant estimates. We attempt to address this concern by showing that our estimates are robust to a wide array of specification of standard errors (see Table 1.16). We also run a randomization check, which yields estimates in line with the p-values from our regressions. Additionally, as shown in Figure 1.6 and Figure 1.13, our regressions also exploit the precise timing of outages. If many zeros would lead us to mechanically find significant effects, this should also be observable in the week before and after the outages, which is not the case in the data. Taken together, the sum of these tests show no evidence that we are overly likely to reject the null hypothesis.

A second concern is that the linearity of the conditional expectation function is violated in our setting and that our coefficients thus misleading. While this seems unlikely as the effect of internet outages is even visible in the raw data Figure 1.5b. To mitigate this concern, we re-code the main interaction of interest as the interaction of two indicator variables. In particular, we define periods of "high sentiment" as those in in the top 50% of the weekly number of posts about refugees. We create a dummy for "High Exposure" for towns in the top 50% of the AfD Facebook users to population ratio.

We then estimate our baseline regression models using "dummified" interaction terms.[31] The resulting saturated model is completely general and yields unbiased estimates without assumptions about functional form (Wooldridge, 2001, p.456-457): the dummies pick up the mean difference in the number of anti-refugee incidents when refugee salience and exposure to it are high. In fact, these models fit the conditional expectation function perfectly no matter how the dependent variable is distributed (Angrist and Pischke, 2008, p.38). As a result this approach naturally also accommodates rare events. As we report in Table 1.27, our baseline results are essentially unchanged in this specification.

To reiterate, we find a correlation between refugee incidents and posts on the AfD page in the time series; a link between these incidents and the usage of the AfD page across

[31] We also create dummies that split each control variable at the median and interact these with the refugee post measure as well as the remaining terms when when including outage interactions.

towns; a panel correlation using the variation in both; and an event study effect of Facebook and internet outages. The panel results also hold in a highly restricted panel where refugee attacks are not rare events and in a saturated model that picks up mean differences without assumptions about functional form. Given the range of approaches, we conclude that it is highly unlikely that another form of rare events bias drives our findings.

Table 1.27: Fully Saturated Models

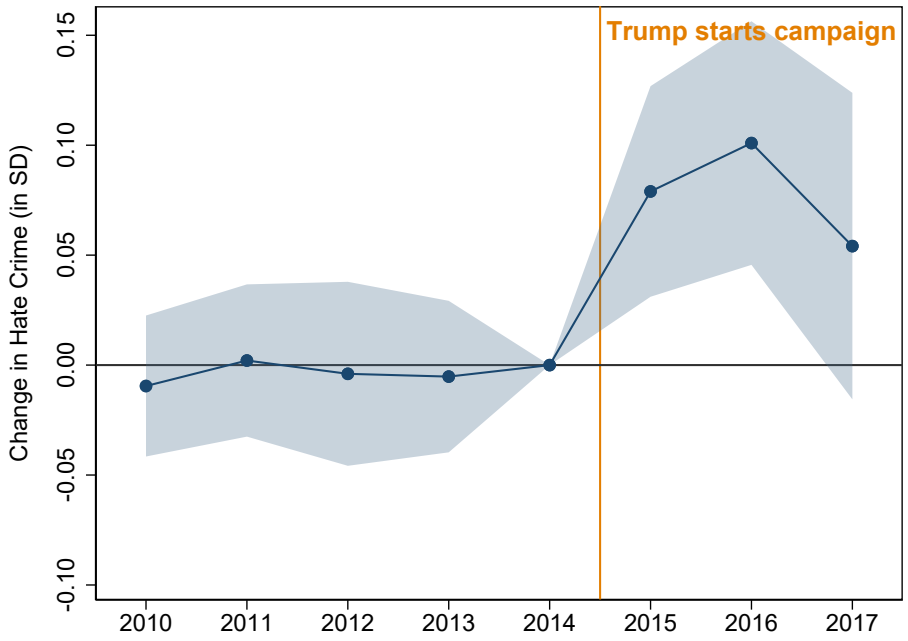
	(1) Panel interaction	(2) Internet outages	(3) Facebook outages
High Exposure × High Sentiment	0.002***	0.002***	0.002***
	(0.000)	(0.000)	(0.000)
Outage		-0.004*	
		(0.002)	
High Exposure × High Sentiment × Outage		-0.009***	-0.003*
		(0.003)	(0.002)
Observations	474,303	474,303	474,303
R-squared	0.082	0.082	0.082
Municipalities	4273	4273	4273
Total attacks	2681	2681	2681
Mean attacks	0.006	0.006	0.006
Municipality FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
All controls [8] × Posts	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). *High Exposure* is a dummy equal to 1 for towns in the top 50% of the *AfD users/Pop.* ratio. *High Sentiment* is a dummy equal to 1 for weeks in the top 50% of *refugee posts*, the Germany-wide number of posts on the AfD's Facebook wall containing the word refugee ("Flüchtling"). The baseline control variables are interacted with the $Refugee\ posts$ measure; see text for a description of the controls. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Table 1.28: Addressing Many Zeros in the Dependent Variable

	Panel A: Baseline			
	(1)	(2)	(3)	(4)
	$\geq$ 1 attacks	$\geq$ 2 attacks	$\geq$ 3 attacks	$\geq$ 4 attacks
Refugee posts × AfD users/Pop.	0.194***	0.276***	0.243*	0.305
	(0.051)	(0.092)	(0.138)	(0.203)
	Panel B: Internet outage only			
	(1)	(2)	(3)	(4)
	$\geq$ 1 attacks	$\geq$ 2 attacks	$\geq$ 3 attacks	$\geq$ 4 attacks
Outage	-0.013***	-0.032***	-0.052***	-0.066***
	(0.004)	(0.009)	(0.015)	(0.019)
	Panel C: Internet outage interaction			
	(1)	(2)	(3)	(4)
	$\geq$ 1 attacks	$\geq$ 2 attacks	$\geq$ 3 attacks	$\geq$ 4 attacks
Refugee posts × AfD users/Pop.	0.194***	0.277***	0.243*	0.306
	(0.051)	(0.092)	(0.138)	(0.204)
Outage	-0.011	-0.035*	-0.043	-0.104
	(0.007)	(0.019)	(0.039)	(0.065)
AfD users/Pop. × Posts × Outage	-0.657***	-1.518***	-1.776***	-3.597***
	(0.205)	(0.311)	(0.492)	(1.008)
Observations	136,641	62,715	32,856	20,535
R-squared	0.074	0.090	0.107	0.122
Municipalities	1231	565	296	185
Total attacks	2848	2182	1677	1375
Mean attacks	0.021	0.035	0.051	0.067
Municipality FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes

Notes: This table presents the estimated coefficients from a regression of hate crimes against refugees on the interaction of local social media usage and anti-refugee sentiment as in Equation (1.1). In each column, we successively restrict the sample to municipalities with a total of at least 1, 2, 3 or 4 attacks on refugees over the sample period. Robust standard errors in all specifications are clustered by municipality. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.



2) From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment

Carlo Schwarz (University of Warwick)

Karsten Müller (Princeton University)

2.1 Introduction

In this paper, we study whether social media platforms can affect anti-minority sentiments online and offline. We investigate this question in the context of a particularly notable case study: the political rise of Donald Trump. Trump has been widely criticized for his inflammatory rhetoric on Twitter and is frequently cited as an example of how social media can increase anti-minority sentiments (Times, 2017d). Minnesota congresswoman Ilhan Omar, for example, has linked tweets by Trump targeting her Muslim faith to "an increase in direct threats on my life - many directly referring or replying to the president's video" (BBC, 2019).

We interpret Trump's presidential campaign as a shock to the salience of anti-Muslim views, particularly for those exposed to his rhetoric on social media. This interpretation is in line with experimental evidence that Trump's popularity on the campaign trail and subsequent election win increased people's willingness to publicly express xenophobic views (Bursztyn et al., c). Building on this insight, we ask if social media may play a role in propagating of anti-Muslim sentiment and real-life violence.

We start by documenting that the frequency of anti-Muslim hate crimes has doubled since Donald Trump's presidential campaign compared to the presidencies of Barack Obama and George W. Bush. This is particularly striking because Bush's term included a temporary ten-fold increase in such crimes following the 9/11 terror attacks, the largest spike since the beginning of the FBI records in 1990 (Gould and Klor, 2016; Panagopoulos, 2006; Hanes and Machin, 2014). It is also consistent with evidence that the Muslim community has been particularly affected by Trump's political rise (e.g. Hobbs and Lajevardi, 2019).

We investigate the potential role of social media in enabling such hate crimes using a difference-in-differences approach. We find that the increase in hate crimes targeting Muslims predominantly originates in counties with high Twitter usage. We also observe disproportionate increases in tweets containing the hashtags #BanIslam and #StopIslam in these counties. These regressions, however, may not isolate a pure "social media effect" because counties with many Twitter users likely also differ in many unobservable dimensions. This may bias our estimates upwards or downwards, depending on how individuals select into social media usage. For example, areas where many people use relatively new technologies such as Twitter may react less because they are more liberal and tolerant, which could bias our estimates downwards. On the other hand, such areas may have a larger share of minority groups and thus more potential targets for perpetrators of hate crimes.

To overcome these concerns, we construct an instrument for county-level Twitter usage in the United States based on the home towns of the platform's early adopters at the South by Southwest (SXSW) Festival in March 2007.[32] SXSW is widely regarded as the tipping point for Twitter's popularity and an important early catalyst for the site's success. One indication of SXSW's importance in explaining Twitter's trajectory is that the number of daily tweets *tripled* during the festival. We also find that tweets about SXSW are a clear outlier in 2007 compared to those about other, considerably more popular festivals, such as Burning Man, Coachella or Lollapalooza. We show that activity on Twitter grew rapidly in the weeks following SXSW 2007, and disproportionately so in the home counties of SXSW followers who signed up in March 2007.

In line with the literature on path dependence in technology adoption (e.g. Arthur, 1989, 1994; Liebowitz and Margolis, 1999; Arrow, 2000), this early expansion left its imprint on the geographical distribution of social media usage in the United States. The locations of Twitter's early adopters at SXSW are a strong predictor of county-level Twitter usage today, even after controlling for the locations of SXSW followers that had already signed up prior to the festival. This result is also robust to using alternative control sets, e.g. using the locations of Twitter users mentioning other major festivals in 2007 or those tweeting about SXSW before the 2007 event. Similar to the strategy of Enikolopov et al. (2020), the identifying assumption is that differences in the locations of SXSW followers in March 2007 relative to earlier months are not related to unobserved county characteristics that explain the rise in anti-Muslim sentiment with the 2016 presidential campaign. Because Twitter was largely unknown before SXSW, and these counties do not systematically differ in many observable characteristics, we believe this assumption is credible.

Instrumenting for Twitter usage with SXSW followers in March 2007, we confirm that measures of anti-Muslim sentiments disproportionately increased in areas with higher social media usage. We find that a one standard deviation higher exposure to social media is associated with a 38% larger increase in hate crimes between 2010 and 2017. This increase in hate crimes against Muslims is entirely accounted for by assaults. Exploiting heterogeneity across counties, we further show that most of this effect is driven by areas with higher

[32]SXSW is an annual event, held since 1987, that comprises a number of festivals, conferences, trade shows, and exhibitions. In 2019, more than 230,000 people attended the festivals, where almost 2,000 acts from all over the world performed. More than 70,000 people attended the SXSW conference, which featured almost 4,800 speakers. Around 30,000 people attended SXSW Interactive, which focuses on emerging technology. For simplicity, we refer to the event as "SXSW festival" or similar short forms throughout the paper.

pre-existing anti-minority bias. These findings suggest that social media platforms may have played a role in the recent spread of anti-Muslim sentiment in the United States by reinforcing existing tensions.

We also find a similar but slightly weaker pattern for hate crimes targeting Hispanics, the second minority group often targeted by Trump. While data from the FBI suggest that the frequency of these incidents has been largely unchanged, our results point to a potential role of social media in contributing to a geographical reallocation of these crimes.

To determine if Trump's tweets contributed to the increase of anti-Muslim sentiment on Twitter, we analyze Trump's Twitter feed. We find a strong time series correlation between Trump's tweets on Islam-related topics and the number of anti-Muslim hate crimes after the start of his presidential campaign, even after controlling for general attention paid to topics associated with Muslims. There is no correlation between Trump's tweets and hate crimes with other motives (e.g. racial hate crime), which suggests that we are not merely capturing waves of general anti-minority sentiment. We also find no such link for the period before the time of Trump's presidential campaign.

To establish causality, we leverage Trump's well-documented golf habit. This analysis is motivated by the fact that many commentators have argued that golfing shifts Trump's state of mind. In 2017 alone, Trump played golf on more than 90 days. In the data, we find a clear pattern: Trump's golf days coincide strongly with changes in the content, but not the number of his tweets. In particular, Trump is more likely to send messages aimed at Muslims and the media on his golf days, and fewer about policy, a fact we exploit in an instrumental variable framework. One intuitive explanation of this finding is that day-to-day politics may be less salient to the President when outside of Washington, DC. Additionally, there is anecdotal evidence that Trump may be influenced by his social media director Dan Scavino – former manager of Trump National Golf Club Westchester and Trump's former caddie – who has been linked to particularly inflammatory tweets (New York Times, 2018).

Using golf days as an instrument, we find evidence consistent with the idea that Trump's tweets about Muslims "trigger" waves of anti-Muslim sentiment. In particular, we find that his instrumented tweets not only continue to predict the frequency of hate crimes, but also measures of media attention paid to Muslim-related topics. Using transcript data on the reporting of the major cable news networks Fox News, CNN, and MSNBC, we show a time series correlation between Trump's golf-induced tweets and mentions of Muslims. This link seems to be largely driven by Fox News, which tends to support rather than oppose Trump's

rhetoric. Analyzing over 100 million tweets, we also find that Trump's anti-Muslim tweets are widely shared by his followers, who further produce their own anti-Muslim content.

Additionally, we investigate whether the transmission effects of Donald Trump's tweets are stronger in counties with more Twitter users in a panel regression setting. Interacting county-level Twitter usage and Trump's Twitter activity, we document that the spike in anti-Muslim hate crime in the days after Donald Trump's tweets is driven by counties with higher Twitter penetration. These findings also persist when we estimate regressions in reduced form and two-stage least squares using our SXSW instrumental variable strategy.

Taken together, our evidence is consistent with the interpretation that, with the start of Donald Trump's presidential campaign, social media may have come to play a role in the increase of anti-Muslim sentiments in the United States. The existing literature broadly suggests three possible mechanisms to explain our findings: coordination capabilities, persuasion, and changes in social norms. We discuss how our findings line up with these three mechanism at the end of the paper. While all are likely at play, some of our results suggest that social media may influence the perception of which beliefs about minorities are socially acceptable. In other words, social media could have enabled changes in social norms for people at the fringes of the political spectrum. Because Twitter users are predominately male and more ideologically extreme than the general population (Barberá and Rivero, 2015), this may explain how social media can contribute to an increase in hate crimes.[33]

Our paper contributes to the literature on the relationship between media consumption and violence. Yanagizawa-Drott (2014), Adena et al. (2015), and DellaVigna et al. (2014) find that traditional media can contribute to ethnic hatred and violence. Other research has linked media such as television (Card and Dahl, 2011) and movies (Dahl and DellaVigna, 2009) to short-lived spikes (or decreases) in violence. Bhuller et al. (2013) document increases in sex crime associated with the roll-out of broadband internet in Norway; Chan et al. (2016) find a correlation between broadband availability and hate crimes in the US. Our findings speak to the role of social media in the spread of violence against minority groups.

We most directly contribute to a growing literature on the influence of social media on real life outcomes. Enikolopov et al. (2020) show that social media can increase participation in protests in Russia by reducing coordination costs. Petrova et al. (2017) study whether

[33]These findings are also consistent with studies on the demographics of social media consumption. Guess et al. (2018) and Guess (2018), for example, show that consumption of fake news articles and ideologically extreme content is driven by relatively few people, which might overlap with the few potential perpetrators of hate crimes.

adopting Twitter helps politicians attract donations. In previous work, we found evidence that social media affects the propagation of anti-refugee incidents in Germany, using Facebook and internet disruptions as a source of short-lived exogenous variation (Müller and Schwarz, 2018a). Here, we study the medium-term effects of social media and highlight a potential social norms channel, based on the particularly salient case study of Trump's presidency.

A separate related literature studies political polarization. While there is evidence that polarization has increased over the past decades (Fiorina and Abrams, 2008; Gentzkow, 2016; Draca and Schwarz, 2018), existing studies have found no or even a negative correlation with social media use (Boxell et al., 2017; Barberá, 2014).[34] One interpretation of our findings is that social media may not necessarily affect *average* outcomes, but rather enable those with extreme viewpoints to find sources of social legitimacy. A widely shared discriminatory tweet by the President, for example, could signal to potential perpetrators of hate crimes that their actions are more widely accepted than they really are.

In Section 2.2, we introduce the data sources and present descriptive evidence on hate crimes since 1990. In Section 2.3, we discuss our empirical strategy and introduce our instrument for Twitter usage based on the SXSW festival. Section 2.4 presents the main empirical results. In Section 2.5 we discuss evidence for the link between Trump's tweets and anti-Muslim sentiment. In Section 2.6 we show that the relationship between Trump's tweets and anti-Muslim hate crime is driven by counties with high Twitter usage. Section 2.7 discusses plausible mechanisms behind our results and potential reporting biases. Section 2.8 concludes.

2.2 Data and Background

We create two datasets for our analysis. First, we build a county-level dataset for the US containing information on hate crimes, Twitter usage, and numerous other variables. Second, we construct a daily time series dataset that combines Trump's daily Twitter activity, the number of total hate crime incidents in the US, data on TV news coverage, and time series control variables. The key sources we draw on are (1) hate crime data reported by the FBI's Uniform Crime Reporting (UCR) program; (2) a county-level measure of Twitter usage based on 475 million tweets collected by Kinder-Kurlanda et al. (2017); (3) hand-collected

[34] A separate literature has analyzed the effects of the media on elections and other political outcomes. See, among others, the work by Adena et al. (2015), DellaVigna et al. (2014), Stephens-Davidowitz (2014), Gavazza et al. (2018), Gentzkow (2016), and Martin and Yurukoglu (2017).

county-level data on the locations of early adopters of Twitter in 2006 and 2007; and (4) information on Trump's golf activity from his inauguration in early 2017 until the end of that year. We describe these and all other data sources in more detail in the following subsections. Table 2.18 and Table 2.19 in the online appendix present the full descriptive statistics.

2.2.1 FBI Hate Crime Data

The data on hate crime in the US come from the FBI and are available for the years 1990 until 2017.[35] The data set contains all hate crimes in the US that are reported to the FBI as part of the Uniform Crime Reporting (UCR) program. The FBI defines a hate crime as:

> "[...] criminal offenses that are motivated, in whole or in part, by an offender's bias against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity." (FBI, 2015, p. 4)

To classify hate crimes, the FBI uses a two-tier decision making process. First, the law enforcement officer recording an incident has to decide whether it might constitute a hate crime. Second, the potential hate crime cases are forwarded to and evaluated by officers with special training in hate crime matters. The FBI (2015) states (p. 35): "For an incident to be reported as a hate crime, sufficient objective facts must be present to lead a reasonable and prudent person to conclude that the offender's actions were motivated, in whole or in part, by bias." For more information on the FBI classification procedure see appendix 2.9.1

Because considerable evidence needs to be available for an offense to be classified as a hate crime, the numbers reported by the FBI have been criticized as underestimates (ProPublica, 2017; News, 2017).[36] Nonetheless, the FBI data constitute the most complete record of hate crimes committed in the United States for which incident details are available. Among others, they include information on the exact date of the crime, the type of crime (e.g. vandalism, theft, assault), the number of victims, and the number of perpetrators. The data further make it possible to assign hate crimes to counties using the county location of the more than 32,000 original reporting agencies based on their Originating Agency Identifier

[35] Note that data for the year 2018 will only become available in November 2019.

[36] Note that time-invariant reporting bias across counties is unlikely to drive our results. First, the US-wide trend of hate crimes reported to the FBI is likely to be highly correlated with the "true" hate crimes trend. Second, we accommodate potential geographical reporting differences in our cross-sectional tests by estimating our model in first-differences. In further robustness checks we restrict the sample to counties where at least one hate crime is reported. We discuss the extent to which changes in reporting over time may explain our results in the results section.

(ORI).[37] Figure 2.2a plots the geographic distribution of hate crimes across the mainland USA.[38] The counties in grey never report any hate crime to the FBI.

The FBI differentiates hate crimes by motivating bias (e.g. anti-Muslim). Overall, they report 34 bias motivations for the broad categories race, religion, sexual orientation, disability, and gender/gender identity. We report all codes for the motivating bias in Table 2.12. We use this classification to identify hate crimes against Muslims. The other categories used in the paper are defined according to the codes listed in Table 2.11.

Presidents and Trends in Hate Crimes To motivate our analysis, we begin by investigating how the number of hate crime incidents has evolved over time. In particular, we test for changes in anti-Muslim hate crimes since the commencement of Trump's presidential run. Panel A of Figure 2.1 plots the average number of weekly anti-Muslim hate crimes for each president since George H. W. Bush; we also plot the 95% confidence interval around the mean.[39]

We split the presidency of Barack Obama into two periods based on Trump's official campaign start. We use this time split because Trump's presidential run does not only mark a cesura for Trump's presence in the media, but is also an important breaking point in his Twitter reach. Figure 2.3a shows that the number of retweets Trump received grew considerably with each month of his presidential campaign.

Over the 27-year period for which the FBI publishes data, the number of hate crimes against Muslims in the United States has increased. Anti-Muslim hate crimes were somewhat less common under Obama than under George W. Bush. Most strikingly, the period after Trump's presidential campaign commenced is a clear outlier by historical standards: the average number of anti-Muslim hate crimes doubled compared to Obama's presidency before Trump's campaign. This increase still stands out in comparison to George W. Bush's presidency, which included the largest recorded spike in anti-Muslim hate crimes in the wake of the 9/11 terror attacks (Gould and Klor, 2016; Panagopoulos, 2006; Hanes and Machin, 2014).

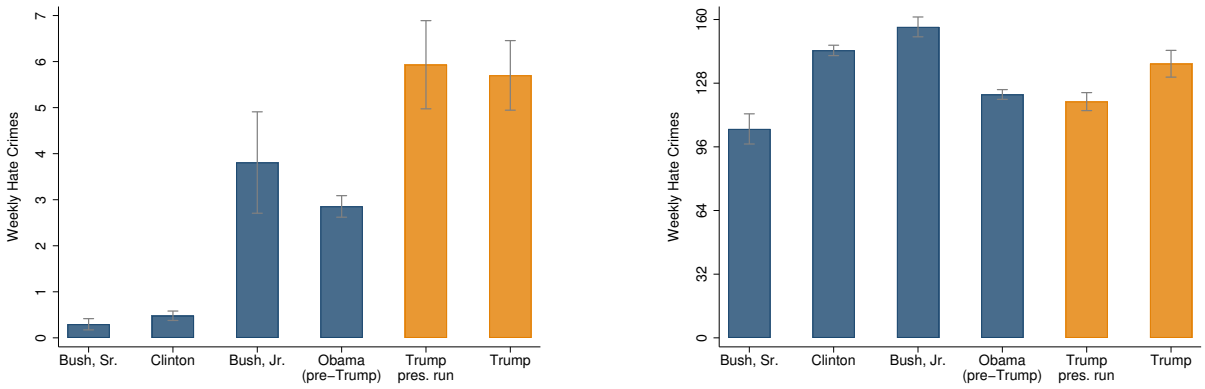
[37] In the rare cases where an agency is located in more than one county we assign the hate crime to all counties the agency is active in; this only applies to 0.08% of all incidents.

[38] The FBI hate crime data do not contain information on the US territories of Virgin Island, Puerto Rico, Northern Mariana Islands, American Samoa, and Guam.

[39] For Trump's presidency, we only have information until December 31, 2017, since the FBI only publishes hate crime data for the previous year in November. For the presidency of George H. W. Bush we only have data from 1991 onward.

Figure 2.1: Average Weekly Anti-Muslim Hate Crimes Since 1990, by President

(a) Anti-Muslim Hate Crimes **(b) All Hate Crimes**



Notes: This figure plots the average weekly number of hate crimes reported by the FBI, by president. We divide Barack Obama's presidency into the period before and after Donald Trump's campaign start ("Obama (pre-Trump)" and "Trump pres. run", respectively). Panel (a) shows the number of anti-Muslim hate crimes. Panel (b) shows the total number of hate crimes. We also plot 95% confidence intervals.

We plot the number of total hate crimes, for which we do not observe a similar increase, in Panel B of Figure 2.1. While we still observe slightly higher numbers compared to Obama, the frequency of hate crimes is lower under Trump than under Clinton or George W. Bush. We show in Section 2.9.2 that this finding also holds true when we split the total number of hate crimes into the underlying categories (e.g. hate crimes motivated by racial bias). We conclude that the beginning of Trump's presidential campaign appears to coincide with a rise in anti-Muslim sentiment in the United States.

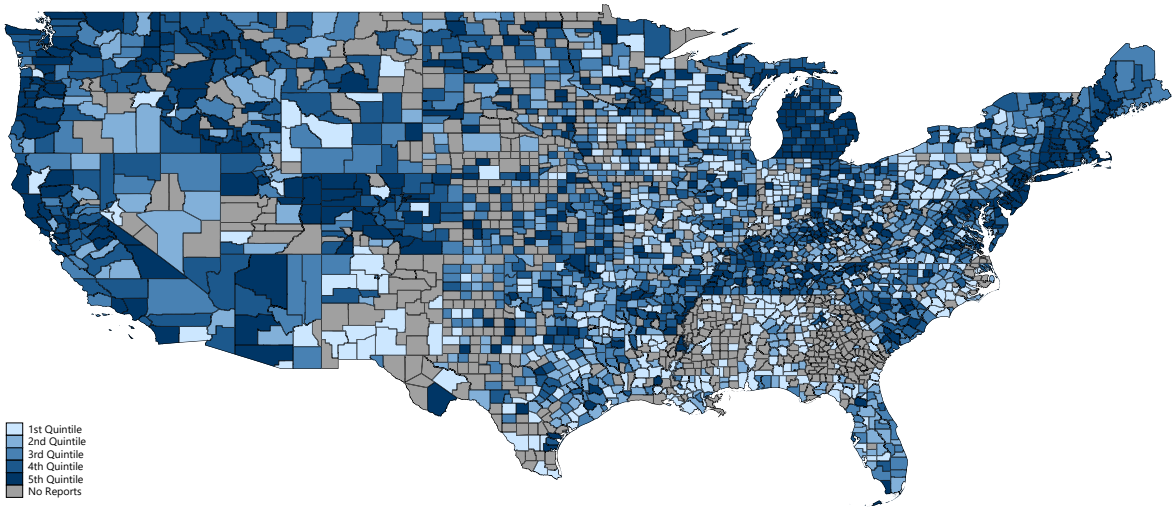
2.2.2 Measuring County-Level Twitter Usage

Twitter does not publish statistics on the number of active users per US county. We create an approximate measure of Twitter usage in each US county using 475 million geo-located tweets collected by Kinder-Kurlanda et al. (2017) made available through the Gesis Datorium. The data were collected between June and November in 2014 and 2015 by repeatedly calling the Twitter streaming API, restricted to US tweets. The streaming API provides a 1% sub-sample of public tweets each time it is called. While the exact underlying sampling procedure is unknown, this process should result in a good approximation of overall Twitter activity.

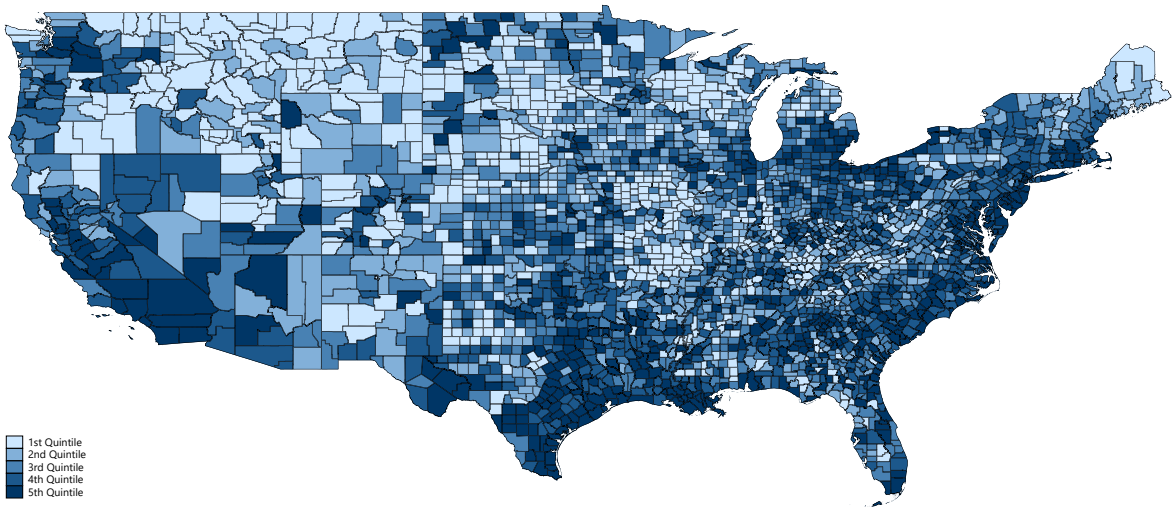
These tweets were assigned to counties based on the geographic location of each tweet. Figure 2.2b visualizes the Twitter activity per capita. Unfortunately, the data do not contain information for Alaska and Hawaii; our analysis therefore focuses on the continental US.

Figure 2.2: Hate Crimes and Twitter Usage by US County

(a) Hate Crimes per Capita



(b) Twitter Usage per Capita



Notes: These maps plot the geographical distribution of the main variables of interest across the counties in the mainland US. Panel (a) plots quintiles of the total number of hate crimes per capita between 1990 and 2017 as reported by the FBI. Counties in grey never reported any hate crime. Panel (b) plots our measure of Twitter usage scaled by population.

2.2.3 Measuring Trump's Twitter Activity

To understand Trump's Twitter activity, we collect the universe of his tweets from the Trump Twitter Archive (Brown, 2018). Our version of this data set contains 35,137 tweets for the time period of April 2009 to November 2018. The data contain the date, time, and text of each tweet and the number of retweets a tweet received.

Identifying Trump's anti-Muslim Tweets We use the text of Trump's tweets to identify tweets about Muslims or Islam-related topics. We start by hand-coding a random subsample of 5000 tweets in which we tag anti-Muslim tweets. These 5000 tweets form the training sample for a machine learning classifier. In preparation for machine learning we remove stopwords from and reduce all words to their morphological routs, so called lemmas. We then extract all unigram, bigrams and trigrams which appear in at least 3 tweets. The extracted n-grams are reweighted using term frequency–inverse document frequency (tf-idf). In this step the the frequency of a n-gram v in document d is replaces by $tfidf(f_{d,v}) = (1 + \ln(f_{d,v}) \cdot (\ln(\frac{1+D}{1+d_v}) + 1)$, where d_v is the number of documents n-gram v appears in. Afterwards, we train a classifier based on a logistic regression model with L1 regularization. We decide the optimal regularization strength using 5-fold cross-validation. The final model achieves and out-of-sample F1 score of 0.97. In the total sample of Trump's tweets the classifier tags 266 anti-Muslim tweets.

As we use the words "muslim", "islam", "terror", "mosque", "refugee", and 'sharia" to collect data on Google searches and news reports on Muslims, we add any tweet containing these words to the set of potential anti-Muslim tweets. This process tags an additional 57 Tweets as anti-Muslim. To rule out that we are picking up unrelated topics by mistake and change the coding of tweets if necessary. In the in the online appendix, we list examples of anti-Muslim tweets (see Table 2.13) and the 25 tweets we removed in the hand-coding step (see Table 2.14).

To further understand the topics of Trump's tweets during his presidency, we use Amazon Mechanical Turk (mTurk) and let three individuals code Trump's tweets in 2017 into the following categories: Media, Islam and Terrorism, Party Politics, Immigration, Foreign Policy, Domestic Policy and Other. We also code the sentiment of each tweet. More specifically, the same three individuals code the sentiment of each tweet either as "very negative", "negative', "neutral", "positive" or "very positive". We recode these categories into a scale from -2 (very

negative) to 2 (very positive). In our analysis we then use the modal topic and the average sentiment coded by the three individuals.

Understanding Trump's Twitter reach. Figure 2.3 shows that Trump has the Twitter reach to potentially influence a considerable fraction of Americans. Figure 2.3a plots the monthly number of retweets he received since joining Twitter. It is apparent that the number of retweets increased with Trump's presidential run (marked by the vertical line). This suggests that a large number of people read his tweets. In Figure 2.14 in the online appendix we additionally show that Trump's tweets about Muslims are significantly more widely shared than his tweets about other topics.

In Figure 2.3b, we plot the number of tweets using the hashtags #StopIslam and #BanIslam, as well as the number of these tweets coming from Trump's Twitter followers (see section 2.2.6). To construct these counts, we obtained the Twitter user IDs of all people who follow Trump on Twitter. The figure shows that the majority of the tweets using these hashtags come from Trump's followers. This lends credence to the idea that many people who harbor anti-Muslim sentiments self-select into following Donald Trump on Twitter, which exposes them to his tweets.

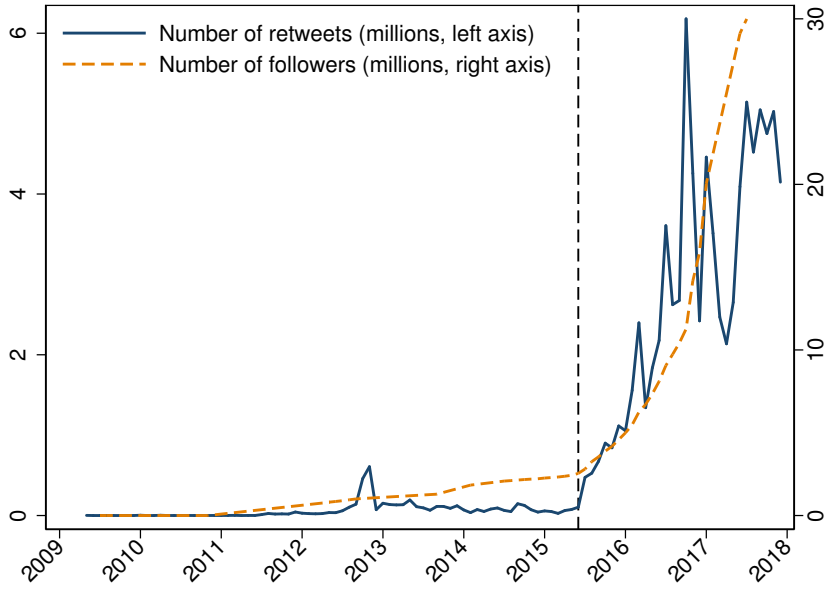
To provide direct evidence for the spillovers of Trump's anti-Muslim tweets on his followers, we collect the tweets for a random 1% sample of Trump's followers. These over 115 million tweets allow us to investigate if Trump's followers react to his content about Muslims.

2.2.4 Twitter Data for South by Southwest and Other Festivals

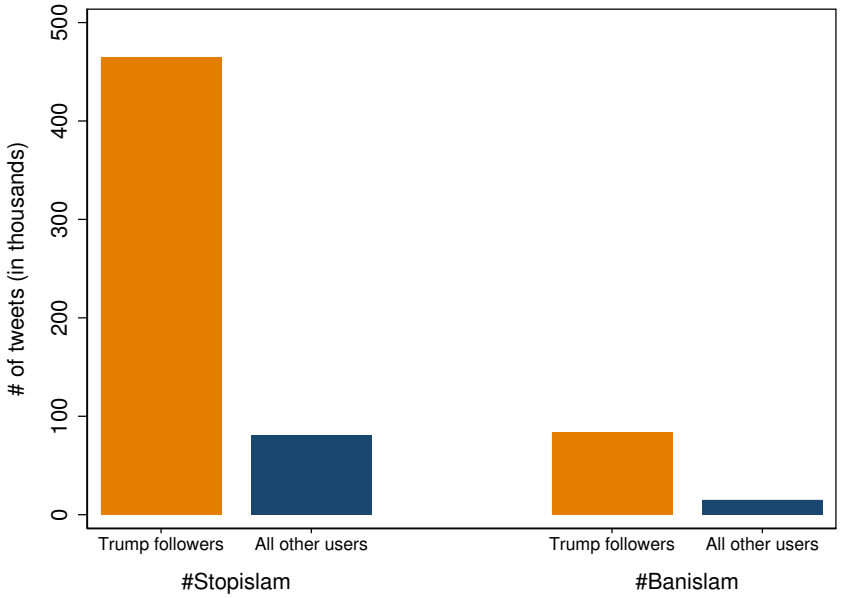
To construct our instrument we collect data using the Twitter application programming interface (API). In particular, we collect the universe of people following the Twitter account of SXSW Conference & Festivals (SXSW). This yields 658,240 unique user IDs. For each of these users, we collect information on their location and the date the account was created. In line with the findings of Takhteyev et al. (2012), around 75% of Twitter users in the sample report their geographical location. Previous research suggests that these user locations yield valid proxies for Twitter usage (e.g. Takhteyev et al., 2012; Haustein and Costas, 2014). As an alternative measure, we also search for tweets containing the term "SXSW" in the year 2007. We do not search for hashtags, since Twitter only formally adopted these in July 2009. In total, we find 5,933 tweets mentioning the SXSW festival.

Figure 2.3: Trump's Twitter Reach

(a) Trump's Retweets Over Time



(b) Trump Followers and Anti-Muslim Tweets



Notes: Panel (a) plots the number of monthly retweets (in millions) Trump's Twitter account received since he joined the site in 2009. The vertical line marks the start of his presidential campaign in June 2015. Panel (b) plots the number of tweets containing the hashtags #StopIslam or #BanIslam sent between 2010 and 2017, which we interpret as clearly expressing negative sentiment towards Muslims. The orange proportion of the bar indicates the number of these tweets posted by followers of Trump's Twitter account.

To compare Twitter activity at the 2007 SXSW festival to other festivals in the same year, we additionally collect the tweets and user data for the Austin City Limited Festival, Burning Man, Coachella, Electric Daisy Festival, New Orleans Jazz and Heritage Festival, Lollapalooza, Pitchfork Music Festival and the West by Southwest Festival. The full list of search terms for these festivals can be found in Table 2.15.

Since we are also interested in the impact of the SXSW festival on overall Twitter activity, we create a proxy for the total number of tweets using the 100 most common English words for January through March 2007 (the full list of words is reported in Table 2.16). While this approach does not give us the universe of tweets in this time window, it should serve as a valid proxy for how many people are using Twitter over time.

2.2.5 Information on Trump's Golf Trips

Information on Trump's golf outings was collected by the New York Times (NYT, 2019). The information covers Trump's travels and identifies sources indicating that he was in fact golfing on any given trip. We cross-check these data using information from *trumpgolfcount.com* and the official Presidential schedule from the White House. In this process we add a few additional days of golf. Table 2.17 in the online appendix describes these sources in more detail; Figure 2.23 graphs the days in 2017 Trump spent golfing, where the darker shade of orange indicates golf outings longer than three days. More than two thirds of golf days are on the weekend, although he has also golfed multiple times on all days of the week (also see Table 2.32 in the online appendix).

2.2.6 Additional Data Sources

We construct a large number of additional variables, which mostly serve as controls. A more detailed variable description and the relevant data sources can be found in Table 2.9.

County-level variables We collect demographic control variables at the county level from the United States Census and the American Community Survey. In particular, we use information on the yearly population, the share of the population by age group, the ethnic composition of the population, the poverty rate and education levels. Information on a county's unemployment rate and industry level employment shares were obtained from the Bureau of Labor Statistics. County-level election results are available from the webpage of

the MIT Election lab. The number of Muslims in each US county is derived from the 2010 US Religious Census. Additionally, we make use of county-level crime statistics based on the FBI's UCR data. Information on TV viewership patterns was collected from Simply Analytics.

We create proxies for anti-Muslim Twitter content by collecting tweets containing the hashtags "#BanIslam" or "#StopIslam" from 2010 to 2017. We selected these hashtags because they are both clearly anti-Muslim and commonly used on Twitter (Miller and Smith, 2017). Following the same procedure as for the SXSW tweets, we assign these tweets to counties based on the location of the users.

Lastly, we study potential preexisting prejudices and xenophobic sentiments at the county level based on data on hate groups from the webpage of the Southern Poverty Law Center (SPLC). The data contain information on the name of the state and city a hate group is active in. We use this information to assign the hate groups to counties. While the classification of hate groups is subjective and subject to controversy, the information gathered by the SPLC is widely used as a proxy for where hate groups are located.[40]

Time series variables To study the content of cable news, we collect TV news mentions of Muslims from the TV News Archive of the Internet Archive. We scrape news mentions for Fox News, CNN and MSNBC based on the same search terms we used for the initial classification of Trump's tweets ("sharia", "refugee","mosque", "muslim", "islam"). In total we collect 82,520 news mentions from the start of Trump's presidential campaign to the end of 2017.

We are also interested in the overall salience of Islam-related topics on the internet. We use Google Trends to obtain daily trends for the above search terms for the US. Unfortunately, Google trends only allows us to collect the daily search interest for a 90 day period. We therefore separately collect the Google trends in 90 day intervals for the period since Trump's presidential campaign commenced. Since Google normalizes the search interest between 0-100 for each 90 day period, we use the weekly search interest, which is available for the period as a whole to bring the daily search to the same scale. We describe this process in more detail in Section 2.9.1.

[40]Note that, as long as the geography of potential misclassification of hate groups by SPLC is random, this will bias our estimates towards zero.

Lastly, we compile information on terror attacks by Islamist from the Global Terrorism Database. In particular, we calculate the daily number of Islamist terror attacks. We split terror attacks by their location and consider terror attacks that occur in the US, Europe, or other locations separately. For the years 2015-2017 our data contain 182 terror attacks.

2.3 Social Media and Anti-Muslim Sentiment

2.3.1 Introductory Correlations

Could social media play a role in the spread of anti-Muslim sentiments starting around the time of the 2016 presidential campaign? If that were the case, we would expect the increase in hate crimes documented in Figure 2.1 to be concentrated in areas where many people use Twitter. To get a first pass at this question, we estimate panel regressions in the following form:

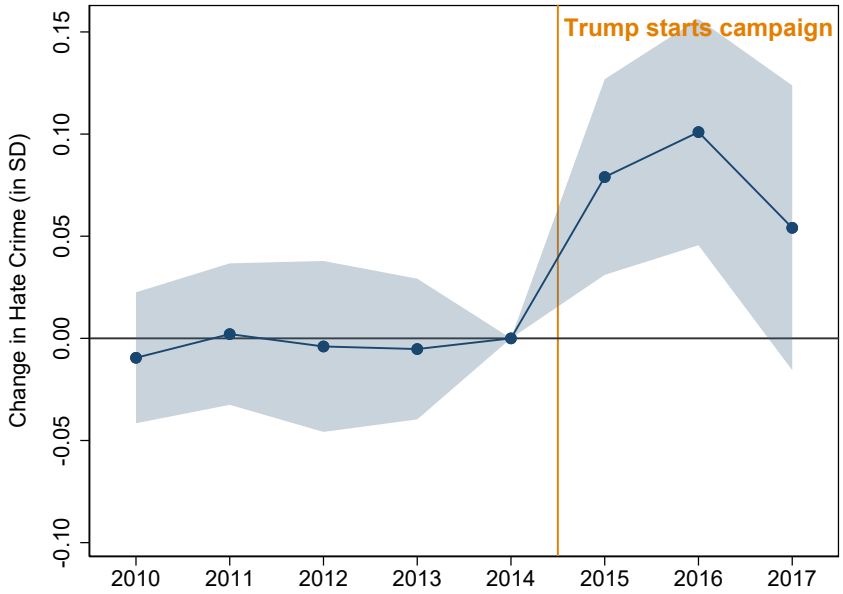
$$\begin{aligned} Hate\ Crimes_{cw} = &\sum_{y=2010}^{2017} \beta_{\tau=y} \cdot Twitter\ Usage_c + \mathbf{X'_{cw}}\gamma \\ &+ County\ FE + Week\ FE + \epsilon_{cw} \end{aligned} \tag{2.3}$$

where the outcome variable is the natural logarithm of anti-Muslim hate crimes in county c and week w (with one added inside). $Twitter\ Usage$ is the natural logarithm of the total number of tweets in a county (also with one added inside). To simplify the interpretation of the coefficients we standardized the variables to have a mean of zero and standard deviation of one. The county fixed effects in the regression control for underlying differences in the number of hate crimes per county, while week fixed effects absorb changes in such crimes that affect all counties to the same extent. The main regressors of interest are β_τ, which measure the differential change in anti-Muslim hate crimes in counties with higher Twitter usage in year τ.

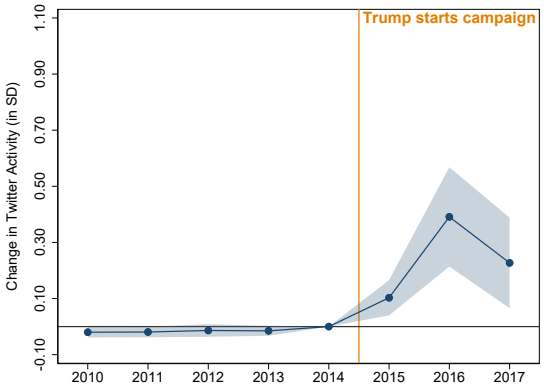
Figure 2.4a plots the estimated coefficients of Equation (2.3). The figure reveals that the increase in anti-Muslim hate crimes starting in 2015 appears to be concentrated in areas with high Twitter usage. The coefficients for previous years are close to zero and not significant, which suggests the counties followed similar trends in the pre-period. Given that all coefficients have been standardized the magnitude of the coefficients indicate that a one

Figure 2.4: Twitter Usage and the Increase in Anti-Muslim Sentiments

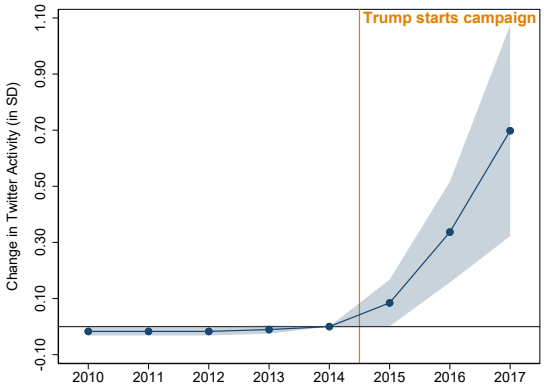
(a) Anti-Muslim Hate Crimes



(b) Tweets Containing #StopIslam



(c) Tweets Containing #BansIslam



Notes: These figures plot the coefficients from running event study regressions as in Equation (2.3). The dependent variables are the natural logarithm of anti-Muslim hate crimes in panel (a) and the number of posts containing #StopIslam and #BanIslam in panels (b) and (c). We standardized the variables to have a mean of zero and standard deviation of one. The omitted category is the year leading up to Trump's presidential run. The vertical line indicates the approximate start of Trump's presidential campaign in June 2015. The shaded area indicates 95% confidence intervals.

standard deviation increase in Twitter usage is associated with an 0.1 standard deviation increase in anti-Muslim hate crime.

As corroborating evidence for the spread of anti-Muslim sentiment via Twitter, we repeat the event study regressions for the hashtags #StopIslam and #BansIslam. Figures 2.4b and 2.4c plots the estimates for these outcome variables. The figures suggest that not only offline but also online sentiments about Muslims grew disproportionately more negative in counties with higher social media penetration.

The evidence here suggests a potential connection between anti-Muslim sentiment and Twitter usage. However, our proxy for Twitter usage is likely correlated with a host of observable and unobservable factors that might also affect hate crimes. To overcome this challenge, in the next section we develop an identification strategy to isolate the effect of social media.

2.3.2 Identification Strategy

The evidence in the previous sections suggests that the increase in anti-Muslim hate crimes around Trump's presidential run has been concentrated in areas with high social media usage. In this section, we address the concern that social media usage may be correlated with other factors by developing an instrumental variable strategy based on the early diffusion of Twitter.

The starting point is a county-level first-difference model relating the shift in anti-Muslim hate crimes in mid-2015 to a measure of social media usage:

$$\Delta Hate\ Crimes_c = \alpha + \beta \cdot Twitter\ Usage_c + \mathbf{X_c'}\gamma + State\ FE + \epsilon_c. \tag{2.4}$$

As a baseline, $\Delta Hate\ Crimes$ will refer to the log-change of hate crime incidents aimed at Muslims or other groups (with one added inside) with Trump's presidential run. The pre-period is defined as the years from 2010 onward.[41] $Twitter\ Usage$ is the natural logarithm of tweets sent from a given county, our measure of social media use. All regressions will control for state fixed effects and dummies for each decile of the population distribution.

$\mathbf{X_c}$ is a vector of control variables that further includes demographic controls for population growth and the share of the population in five-year age buckets; the linear distance from each county centroid from Austin Texas, the location of the SXSW festival we will

[41] In further robustness checks we show that our results neither depend on the pre-period we use in the first-difference nor on the specific functional form. The results also hold for the *level* of hate crimes after Trump's presidential run.

describe in more detail below; controls for ethnic composition and the share of Muslims; socioeconomic controls including the share of high school graduates or people with a graduate degree, the poverty rate, the unemployment rate, local GINI index, the share of uninsured individuals, the log median household income, the employment shares in eight sectors; media controls for the viewership share of Fox News, the cable TV spending to population ratio, and the prime time TV viewership to population ratio; and the county-level vote share of the Republican party in 2012. Standard errors in all specifications are clustered at the state level.[42]

When estimating equation (2) using OLS, the point estimates for β in Equation (2.4) are likely biased because Twitter usage is not exogenous. In particular, one may be concerned that the factors driving people to commit hate crimes are correlated with the decision to adopt social media. This could give rise to alternative interpretations of the graph in Figure 2.4a and the β estimate in Equation (2.4). To give one example, perhaps the potential perpetrators of hate crimes live predominantly in areas with a sizable presence of minority groups, and those areas are also more likely to use Twitter. In that case, the period around Trump's campaign start could still be interpreted as a trigger point for anti-Muslim sentiments, but it is not clear whether or to what extent social media plays a role.

To circumvent this issue, we exploit plausibly exogenous variation in the early adoption of Twitter in the United States. More precisely, we make use of the fact that Twitter's popularity reached a tipping point at the SXSW conference and festival in 2007. During the event, the daily volume of tweets increased from around 20,000 to 60,000 (Gawker, 2007). Figure 2.5a gives a first indication that SXSW may have led to a trend break in the success of Twitter: we see a clear spike of tweets about the event during the SXSW conference in mid-March 2007, followed by an upward shift in the growth of the total number of tweets. While total tweets grew by 60% from February to March, this growth accelerated to over 240% from March to April. March 2007 is also a clear outlier in the number of SXSW followers that signed up to Twitter (see 2.21 in the online appendix).

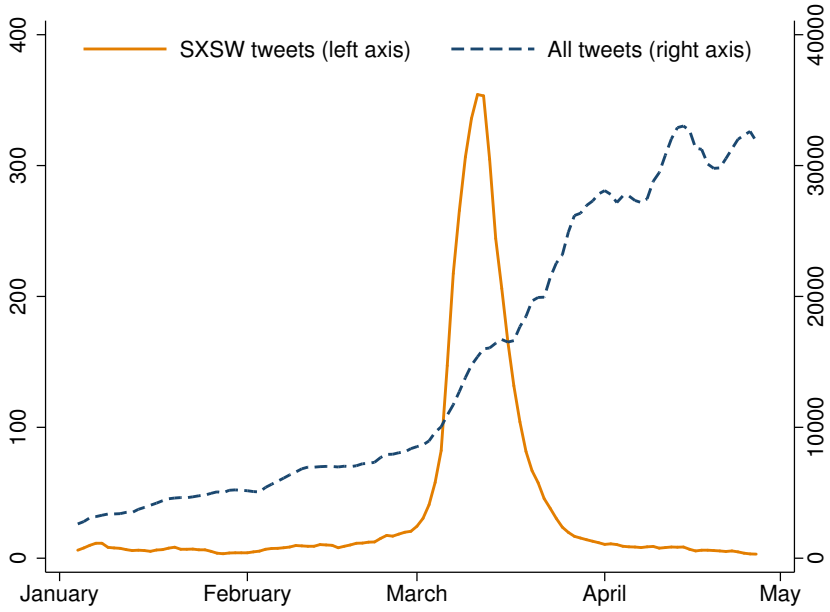
A number of facts suggest that the early adopters at SXSW were key to Twitter's rise. As a first indication, in 2007 there were more tweets about SXSW than about other major festivals (see Figure 2.5b).[43] This is noteworthy because of the lower attendance at SXSW

[42] In Table 2.28 in the online appendix, we show that our results also hold using alternative ways to construct standard errors.

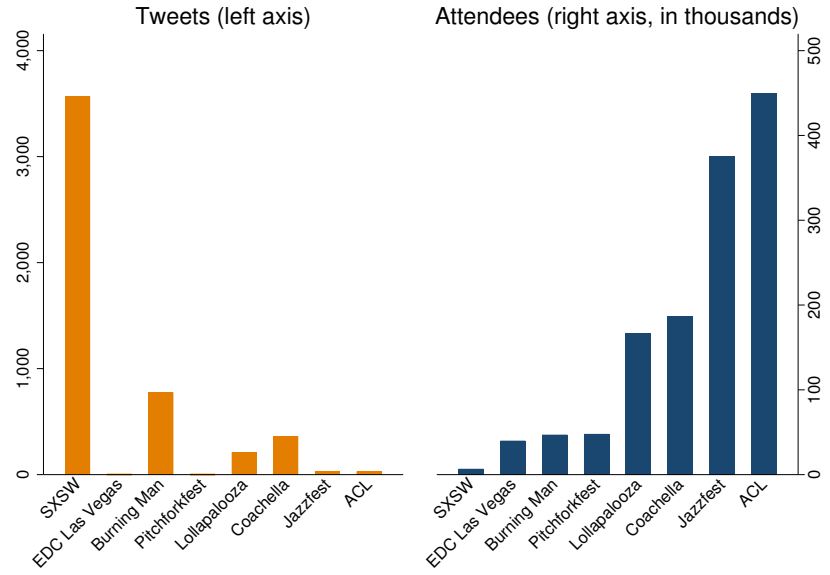
[43] This pattern also holds when we consider tweets about the festivals for the whole of 2007 (see Figure 2.20).

Figure 2.5: South by Southwest (SXSW) 2007 and the Spread of Twitter

(a) Twitter Activity Around SXSW 2007



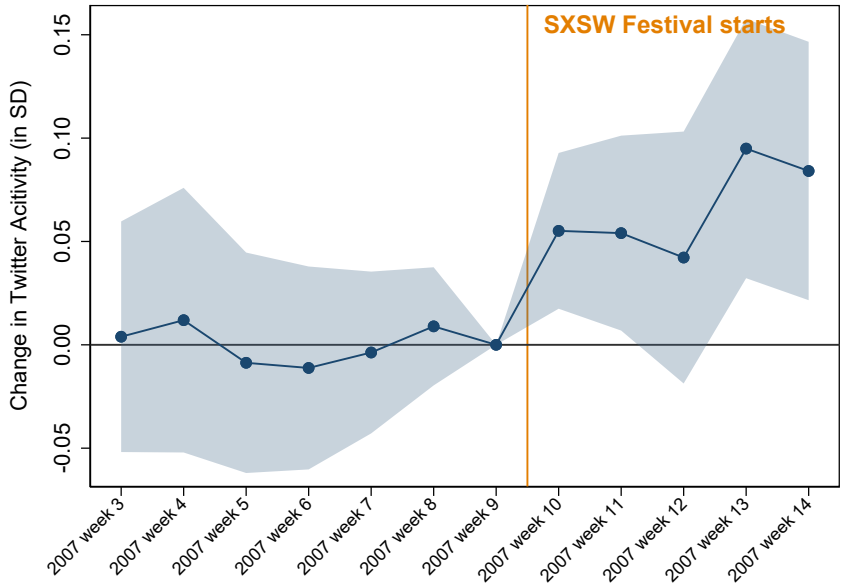
(b) Major Festivals in 2007: Tweets and Attendance



Notes: Panel (a) plot the total number of tweets and the number of tweets containing the term SXSW over time, smoothed using a 7-day moving average. The number of tweets on a given day is based on the 100 most common English words (see Table 2.16). Panel (b) plots the number of tweets mentioning major festivals in 2007 in a 14 day window before and after the event.

Interactive. We can also see that the spread of Twitter across counties followed the early adopters. To show this, we run event study panel regressions to compare Twitter activity in counties with and without new SXSW followers in March 2007. Figure 2.6 plots the results. Areas with early adopters at SXSW did not exhibit a higher growth rate of Twitter activity prior to SXSW Interactive 2007 but the growth rate increased in its aftermath. Quantitatively, counties with a one standard deviation higher number of SXSW followers in March (1.91) increased their local twitter activity by 10% of a standard deviation in April compared to February 2007.

Figure 2.6: The Effect of SXSW on Twitter Adoption



Notes: This figure plots the estimates of β_τ from the panel event study regression $Log(1 + \#\ of\ tweets) = \sum \beta_\tau SXSW\ followers, March\ 2007_c \times Week_\tau + \sum \delta_\tau SXSW\ followers,\ Pre_c \times Week_\tau + County\ FE + Week\ FE + \varepsilon_{cw}$. The number of tweets in a given county and week is based on the 100 most common English words. We standardize the variables to have a mean of zero and standard deviation of one. Standard errors are clustered by state.

We exploit that this pattern of technology adoption persists until today. As we will show below, the number of SXSW followers in a county who registered during the festival period are predictive of Twitter penetration across US counties. This is in line with the literature on the path dependence of technology adoption (e.g. Arthur, 1989, 1994; Liebowitz and Margolis, 1999; Arrow, 2000). Crucially, this is still true after controlling for the number of SXSW followers in a county *prior* to the tipping point in March 2007, or alternatively for users

tweeting about the much more popular festivals Coachella, Burning Man, and Lollapalooza in the same year.

The historical diffusion of Twitter gives rise to a difference-in-difference instrumental variable framework. We collapse the time dimension into an IV setting where the first stage equation is given by:

$$\begin{aligned} Twitter\ Usage_c = \alpha &+ \delta_1 \cdot SXSW\ followers,\ March\ 2007_c \\ &+ \delta_2 \cdot SXSW\ followers,\ Pre_c \\ &+ \mathbf{X'_c}\psi + State\ FE + \xi_c, \end{aligned} \tag{2.5}$$

where $SXSW\ followers,\ March\ 2007$ is the number of SXSW followers in county c that joined Twitter in March 2007, which serves as the excluded instrument. $SXSW\ followers,\ Pre$ are followers that joined before the festival at any point in 2006. This controls allows us to address the concern of inherent differences of counties with SXSW followers.[44]

Similar to Enikolopov et al. (2020), the identifying assumption underlying our empirical strategy is that, conditional on a large number of county characteristics, the decision to start following SXSW in March 2007 rather than in the months before drives increases in anti-Muslim sentiments with the 2016 presidential campaign only through the diffusion of Twitter usage.[45] Three pieces of evidence suggest that this assumption is reasonable. First, as shown above, counties with Twitter adopters around SXSW did not differ in Twitter adoption prior to the festival. This suggests that these counties are not inherently different. Second, a comparison of the Twitter profiles of users signing up for Twitter around SXSW with those who signed up before suggests that they are highly similar. Table 2.21 shows that users' first names and the terms they use to describe themselves are almost indistinguishable between these two groups. The correlation of words mentioned in the "bio" of these groups is 0.92. Third, the home counties of SXSW followers who signed up during the 2007 event do not systematically differ in observable characteristics from those of users who signed up before (see Table 2.20).

Figure 2.13 in the online appendix plots the distribution of our proxy of new SXSW followers in March 2007 across US counties. People from 155 counties were early adopters

[44] In the robustness section below, we consider a large range of alternative control sets based on different time periods to hold selection into social media usage constant.

[45] With the alternative festival controls, the assumption is similar in that attending SXSW rather than other festivals in 2007 should only affect outcomes through this social media adoption channel.

of Twitter at or around the time of SXSW. Table 2.22, also in the online appendix, plots the correlation coefficients between the county-level SXSW measures and those for the other festivals. Although these variables are strongly correlated, as one would expect, there is enough variation in the locations of SXSW followers we can exploit in our empirical strategy. In robustness exercises, we consider a large range of alternative SXSW metrics, some of which show a considerably lower correlation between "treatment" and "control" group.

Since our baseline outcome variable is differenced over time, we also require that the parallel trends assumption holds. We already saw in Figure 2.4a above that hate crimes against Muslims disproportionately increased in areas with high Twitter usage only *after* Trump's presidential campaign started. In the online appendix in Figure 2.16 and Figure 2.19, we provide additional reduced form evidence in support of parallel trends when comparing areas with and without users that likely attended SXSW in March 2007.

2.3.3 South by Southwest and Twitter Adoption: First Stage Results

To assess whether the initial diffusion of Twitter at SXSW still matters for social media use today, we report the results of estimating the first stage Equation (2.5) in Table 2.1. We can see that across the board the number of new Twitter users in March 2007 who followed SXSW is highly predictive of Twitter usage today. The point estimates are always statistically significant at the 1% level. The coefficient for SXSW followers in the months prior to the 2007 event is not statistically significant as soon as we control for observable county characteristics. Indeed, an F-test for the equality of coefficients suggests that the March 2007 and pre-period estimates are also statistically different from each other. Importantly, the coefficient estimates for March are highly stable and do not depend on the included covariates. Quantitatively, the estimate of 0.362 in column 8 implies that a one standard deviation increase in the log number of new SXSW followers in March (0.32) is associated with 12% higher Twitter usage today. The estimated effect based on the pre-period estimate implies 1% more users, which is not distinguishable from zero.

Based on these estimates and the event study plot in 2.6, we conclude that county-level differences in the early adoption of Twitter spread through the 2007 SXSW conference and festival are a reliable predictor of Twitter usage in the United States today. Because the location of early adopters in the period before the festival does not predict Twitter usage, it is

unlikely that this result is driven by selection into following the SXSW festival's Twitter page. In the next sections, we will conduct more robustness checks to test the validity of this insight and will employ the strong first stage result to estimate the effect of social media propagation on the recent rise in anti-minority sentiments, particularly those aimed at Muslims.

2.4 Main Results

2.4.1 Reduced Form Estimates

We next turn to the reduced form estimation results for the change in hate crimes against Muslims around Trump's presidential campaign start. Table 2.2 presents these results. Across a large number of different specifications, we find that the early adoption of Twitter – measured by the number of SXSW followers who joined Twitter in March 2007 – is associated with an increase in hate crimes against Muslims. The estimates for the March coefficient are strikingly similar irrespective of the included control variables. The estimates on new SXSW followers in previous months are not statistically significant and considerably smaller.[46]

Figure 2.16 in the online appendix plots the reduced form estimates from difference-in-difference panel regression of the type in Equation (2.3). Note that this regression also controls for the locations of SXSW followers in previous months interacted with year fixed effects. As above, we find that hate crimes against Muslims did not disproportionately increase in areas with new SXSW followers in March 2007 prior to the period of Trump's presidential campaign. Afterwards, however, these counties experienced a large upward shift in such incidents.

Taken together, we interpret these results as first evidence that social media may play a role in the propagation of hate crimes as a result of Donald Trump's campaign. Because we control for the number of SXSW followers in the months before SXSW 2007, these results are unlikely to be driven by a selection of individuals from areas prone to hate crimes into participation in that particular festival. In the next sections, we provide the formal two stage least squares estimates and conduct further robustness checks in support of this interpretation.

2.4.2 IV Estimates

The results in the previous section can be interpreted as evidence that social media plays a role in the recent increase in hate crimes in the United States. In this section, we use the

[46]Note that the standard deviation of these pre-SXSW users is around half that of the March 2007 variable.

proxy for new SXSW followers in March 2007 as an instrument for Twitter usage across the US today, while holding interest in SXSW prior to the key event constant to alleviate selection concerns.

Table 2.3 provides two sets of results. In panel A, we plot the OLS results from regressions of the change in hate crimes against Muslims on our measure of Twitter usage. In panel B, we report the 2SLS results and a number of diagnostic tests. The results suggest that social media penetration, measured by Twitter usage, is positively associated with the increase in hate crimes against Muslims. The 2SLS estimates in column 8 imply that a one standard deviation increase in Twitter usage (1.91) is associated with a 38% larger increase in hate crimes after Trump's presidential campaign launched.

A well-known concern with IV estimation is the weak instruments problem, which can lead to biased point estimates. We believe that our estimation does not suffer from a weak first stage for three reasons. First, the robust F-statistic for the excluded regressor ranges between 41 and 68 in columns 1 through 8.[47] Second, the values of the F-statistic are above the critical values to reject the null hypothesis of a 5% potential bias with 5% statistical significance derived in Olea and Pflueger (2013), which is 37.42. These authors extend the well-known thresholds of Stock and Yogo (2005) to the case of heteroskedasticity-robust and, relevant in our case, clustered standard errors.

We also assess the significance of our main estimates using confidence sets based on test inversion that are valid whether or not instruments are weak. For the case of a single instrument we study here, Andrews et al. (2019) recommend reporting Anderson-Rubin (AR) confidence sets that are efficient and robust to weak identification (Anderson et al., 1949). Andrews (2018) develops a two-step approach to construct these confidence sets that is implemented in Stata by Sun (2018). Basing inference on this two-step approach sidesteps the issue that the usually reported (Wald) confidence intervals for 2SLS estimates can exhibit large coverage distortions. This is because AR confidence sets allow for inference without assessing the strength of first-stage results in a separate initial step. As such, we can determine whether our second stage coefficients are likely to be non-zero even if our instrument was indeed weak. Reassuringly, the AR confidence sets reported below the (instrumented) Twitter usage in panel B always exclude zero.

[47]Note that because the model is just-identified, the robust F-statistic (sometimes also called Kleibergen-Paap) is equivalent to the effective F-statistic derived in Olea and Pflueger (2013).

Because our estimations do not appear to suffer from a weak instrument problem, we can use the comparison of the OLS and 2SLS estimates to assess whether the selection of individuals into social media adoption is positively or negatively correlated with the incidence of hate crimes. In other words, we can test whether the OLS estimates are upward or downward biased. Across all specifications in Table 2.3, the OLS estimates are highly statistically significant, but considerably smaller than those obtained using 2SLS. This difference suggests negative selection into social media usage. To give one example, if people in particularly xenophobic areas commit more hate crimes but are less likely to use Twitter, the OLS estimate would be downward biased. This selection effect is also consistent with Enikolopov et al. (2020): for the case of social media and protest participation in Russia, they find much larger IV estimates compared to OLS.[48]

In Table 2.27 in the online appendix, we investigate which types of hate crimes increased particularly in areas with higher social media usage. It turns out that our results seem to be almost entirely driven by a rise in assaults. This makes it unlikely that we are capturing changes in *reporting* rather than the actual incidence of hate crimes, since we have no reason to expect reporting changes to be limited to particularly severe cases. We relegate a more extensive discussion of reporting changes to Section 2.7

A conceptual question raised by these estimates is the extent to which any potential causal effect of social media can be directly attributed to Twitter, rather than other platforms. While the initial diffusion through SXSW in 2007 was probably specific to Twitter, there were likely significant spillovers in the adoption of other social media platforms. Since we only observe the equilibrium outcome of these spillovers today, our estimates might not identify a pure "Twitter effect". What matters for the interpretation of our estimates is that this diffusion is limited to social media, which we believe is plausible.

2.4.3 Robustness

We consider a range of sensitivity checks to validate the robustness of our main findings. We begin by reporting an additional set of results that test alternative ways to account for the selection of users into events such as SXSW. In particular, we replace the control variables for new followers of SXSW at any point in 2006 with users tweeting about *other* festivals in 2007 that are, in many respects, very similar to SXSW. We consider tweets about three of

[48] Another interpretation of the 2SLS estimate is that counties with more SXSW followers that signed up in March 2007 have a higher local average treatment effect (LATE).

the most popular festivals in the United States: Coachella, Burning Man, and Lollapalooza. More precisely, we define control variables that capture the log number of users from each county that tweeted about these festivals in the month of 2007 in which they were held.

Table 2.24 in the online appendix reports the results for the reduced form and 2SLS estimations with these alternative controls in panel B and C, respectively. To aid comparison, we again plot the OLS results in panel A. As before, we find that the impact of Twitter usage on changes in anti-Muslim hate crimes is highly statistically significant throughout. Crucially, the log number of users tweeting about the other festivals is statistically insignificant, which is another indication that we are not merely capturing a selection of particular people into areas with hate crimes and high Twitter usage. The estimates and F-statistics for the 2SLS results are somewhat larger than the baseline findings in Table 2.3.

We also consider alternative transformations of the SXSW variables in Table 2.26 in the online appendix. In column 1, we begin by showing that the results also hold when dropping the SXSW control, which makes the results somewhat stronger. In columns 3 through 6, we consider alternative time periods for the pre-period variable or alternatively control for the individual months. Columns 7 through 11 replace the SXSW follower variables with dummies for counties in which we can locate any tweet about SXSW in March 2007 or previous periods. Importantly, these specifications vary widely in the number of "treatment" and "control" counties, as well as the correlation between the treatment and control SXSW variables. Our results are robust throughout, which suggests our findings are not driven by any particular specification.

We also use alternative metrics of Twitter usage in Table 2.25 in the online appendix. We consider two survey measures of Twitter usage provided by GfK Mediamark Research & Intelligence (via SimplyAnalytics), as well as two alternative transformations of the GESIS Twitter data (only tweets before June 2015 or the number of Twitter *users*, rather than the number of tweets). All of these measures yield similar estimates.

In Table 2.4, we present additional robustness checks. In column 1, we drop state fixed effects, which makes little difference to the point estimates. In column 2, we consider the change in anti-Muslim hate crimes since 1990 (rather than 2010); this yields larger estimates throughout. In column 3, we replace the change in hate crimes with the log number of hate crimes after Trump's presidential run as dependent variable. This also yields significant estimates.

In columns 4 through 6 of Table 2.4, we address the concern that anti-Muslim hate

crimes reported by the FBI mainly occur in a relatively small fraction of all counties. In column 4, we begin by dropping all counties that report a zero change in anti-Muslim hate crimes between 2010 and 2017. Because this applies to the majority of counties, the sample size shrinks considerably. One way to think about this estimation is that it captures the intensive margin of hate crimes. Despite the drop in observations, our estimates remain statistically significant. In column 5, we next drop counties for which the FBI always reports zero hate crimes. Reporting may be less reliable for these counties. As it turns out, this exclusion makes little difference for our estimates. As a last exercise, we drop all counties for which the (rounded) estimated share of Muslims in the total population is zero from the sample in column 6.[49] Again, the results we obtain in this sample are very similar to those in the main sample.

In column 7, we weight all estimates by population, which makes little difference to the results. In column 8, we restrict the sample to neighbouring counties where one has no new SXSW followers in March 2007 and the other one has at least one. This is to purge the estimates of potential unobserved local confounders. This yields similar estimates. At last, in column 9, we transform the dependent variable into an index equal to 1 for increases in anti-Muslim hate crimes, 0 for no change, and -1 for decreases; again, our findings remain similar.

2.4.4 Social Media and Changes in Other Hate Crimes

Up to this point, we have focused on changes in anti-Muslim hate crimes, motivated by the fact we found little change in the frequency of other types of hate crimes around the start of Trump's presidential campaign in the FBI data. However, one might expect Trump's presidential run to also affect other categories of hate crimes, in particular anti-Hispanic incidents.[50] If social media plays a role, such incidents may have become more common in areas with high Twitter usage even if their total number remained unchanged.

In Table 2.5, we consider this possibility empirically by replacing the dependent variable with the log change in hate crimes targeting on Hispanic ethnicity, other ethnicities, race,

[49] Although the Religious Census reports no Muslims living in these counties, this might be the artifact of a very small number, rather than an actual zero.

[50] In his presidential campaign announcement speech, Trump infamously singled out Hispanics and Arab Muslims: "When Mexico sends its people, they're not sending their best. ... They're bringing drugs. They're bringing crime. They're rapists. And some, I assume, are good people. ... They're sending us not the right people. It's coming from more than Mexico. It's coming from all over South and Latin America, and it's coming probably – probably – from the Middle East."

sexual orientation or religion (excluding anti-Muslim bias). We also consider hate crime data from the Anti-Defamation League (ADL) as an alternative data source in column 7. The ADL only appear to report a large number of hate crimes from 2016 on, so we focus on the *level* rather than the change in hate crimes.[51]

Overall, we also find a role for social media in explaining increases in the total number of hate crimes and those targeting Hispanics, the other minority group frequently singled out by Donald Trump. However, only anti-Muslim hate crimes show a consistent pattern across the OLS and 2SLS estimates. There is little evidence for a reallocation of other hate crimes towards areas with higher Twitter usage. In the 2SLS estimation, a one standard deviation increase in Twitter usage is associated with a 35% larger increase in total hate crimes, and a 33% larger increase for incidents targeting Hispanics.[52] The difference of these estimates compared to the OLS results likely arises because of selection: social media, and Twitter in particular, is likely adopted more by areas with more technologically-savvy people who are probably less likely to commit hate crimes. This creates a downward bias for the OLS estimates.

2.4.5 Heterogeneous Effects: Social Media and Pre-Existing Bias

The results in the previous sections raise the question whether exposure to social media is changing people's beliefs about Muslims or if social media rather reinforces existing prejudices. To address this question, we investigate whether the effect of Twitter usage is driven by counties that are more likely to be susceptible to anti-Muslim messaging.

In particular, we repeat the event study regressions from Section 2.3.1 and split counties by whether the Southern Poverty Law Center (SPLC) identifies at least one hate group. Note that these sample splits do not estimate whether anti-Muslim hate crimes increased in counties with hate groups but rather whether Twitter usage has a different impact in these counties.

Figure 2.7 plots the estimated coefficients from this exercise.[53] We find that higher Twitter usage is only associated with more anti-Muslim hate crime in counties with hate groups. In contrast, counties with high Twitter usage but no hate group continue to follow the same trajectory as low Twitter usage counties. Quantitatively, among the counties with

[51] In unreported results, we find similar results using a measure of the change in local hate crimes as reported by ADL.

[52] Figure 2.17 and Figure 2.18 in the online appendix plot the OLS and reduced form event study graphs.

[53] To reduce clutter, the figures report the estimated coefficients without confidence bands. We report the full regression results with standard errors in Table 2.29 in the online appendix.

at least one hate group a one standard deviation increase in Twitter usage is associated with a 0.6 standard deviation increase in anti-Muslim hate crime. In Panel (b), we provide similar evidence for counties that are above the 90th percentile of hate crime per capita (all motivating biases) in the pre-period. We again observe that the increase in anti-Muslim hate crimes is driven by counties with high Twitter usage and pre-existing biases.

Taken together, the findings are at least some evidence that social media did not necessarily change people's beliefs, but rather triggered existing negative attitudes towards Muslims around the time Trump started his presidential run. This is consistent with the view that people infer information about the social acceptability of viewpoints and actions based on what they see online. As such, it appears possible that after observing increased anti-Muslim rhetoric on Twitter (as documented above), already radicalized individuals might have become more willing to commit violent acts against Muslims in real life. If this is the case, spikes in anti-Muslim sentiment on social media might work as "triggers", a possibility we investigate in the next section.

It is also worth noting that the sample splits are another indication that we are unlikely to capture changes in the propensity to report hate crimes rather than an actual increase in incidents. We discuss this issue in more detail in Section 2.7.

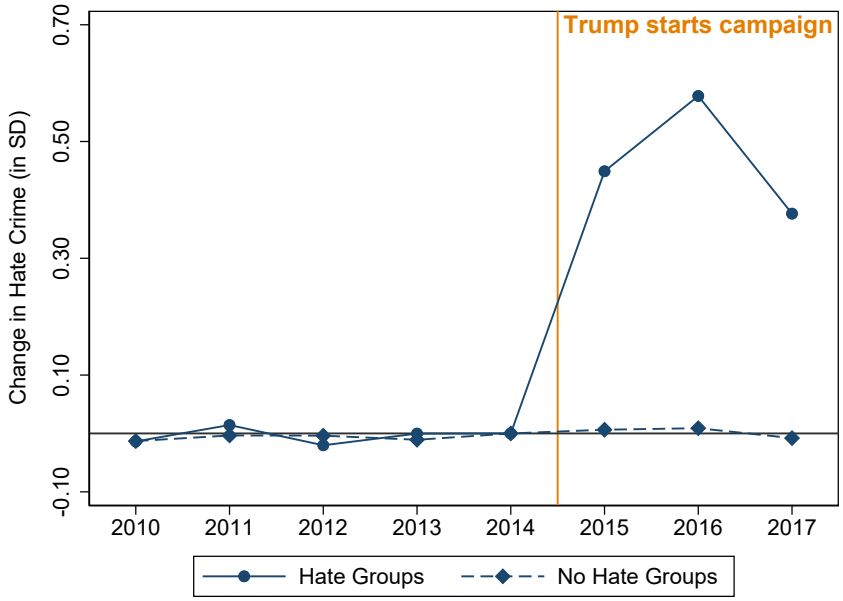
2.5 Trump's Tweets and Anti-Muslim Sentiment

The previous section suggests that social media may have played a role in the spread of anti-Muslim sentiment associated with the start of the Trump campaign. An often-voiced hypothesis is that Trump actively contributes to anti-Muslim sentiment through his incendiary comments on Twitter. Indeed, there is some existing evidence that influential individuals can have a disproportionate effect on public opinion (e.g. Beaman et al., 2009; Bursztyn et al., c; Alatas et al.).

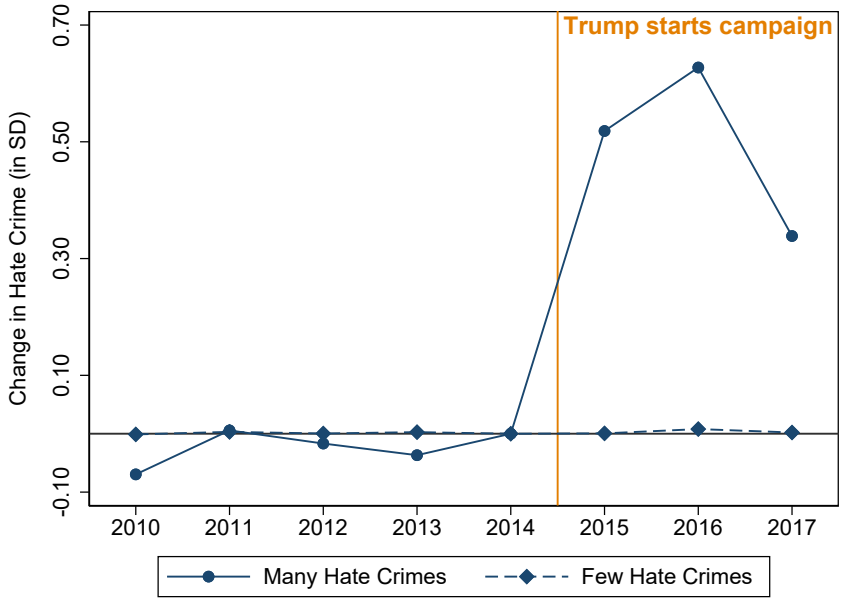
We attempt to shed some light on this mechanism by analyzing the time series relationship between Trump's tweets about Muslims, anti-Muslim hate crimes, and media attention. We attempt to get at the issue of causality by again leveraging an instrumental variable. The main purpose is to provide evidence for a channel through which social media could contribute to a climate that enables hate crimes and investigate the importance of prominent only figures. Table 2.31 and Table 2.37 plot the summary statistics.

Figure 2.7: Heterogenous Effects of Twitter Usage

(a) Split by Existing SPLC Hate Groups Share



(b) Split by Frequency of Hate Crimes in Pre-Period



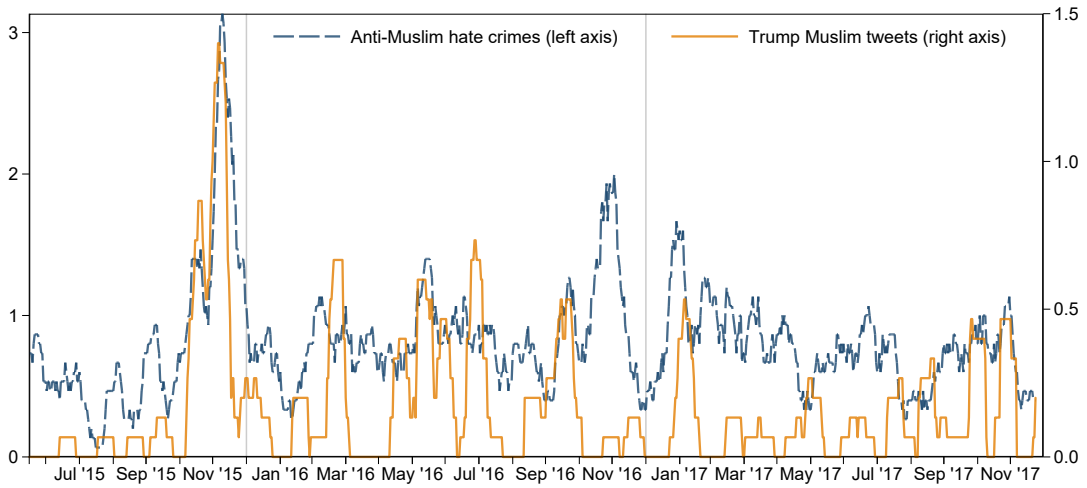
Notes: These figure plot the coefficients of running panel event study regressions as in Equation (2.3). We again standardized the variables to have a mean of zero and standard deviation of one. Equation (2.3) is estimated separately for counties with and without at least one hate group as defined by the Southern Poverty Law Center (SPLC). In panel (b) we split counties at the 90th percentile of the number of hate crimes per capita in the pre-period.

2.5.1 Trump Tweets and Hate Crimes

If there is a relationship between Trump's Twitter activity and physical hate crimes, the timing of both should coincide. We thus begin by plotting the number of Trump's tweets about Islam-related topics and anti-Muslim incidents over time in Figure 2.8. We define these tweets based on a careful reading of Trump's Twitter feed, combined with a machine learning algorithm; see the data section and online appendix Table 2.16 for more details. Since the daily number of tweets is highly volatile, we plot the 14-day moving average of the series; collapsing the data on the weekly level looks very similar (unreported).

It is immediately apparent that Trump's tweets about Muslims and anti-Muslim hate crimes are highly correlated. This correlation could reflect that Trump reacts to US-wide anti-Muslim sentiments driven by observable and unobservable factors, e.g. terrorist attacks. It could also be that Trump's tweets themselves contribute to a climate that enables hate crimes. Clearly, we cannot disentangle these possibilities using the graphical evidence from the data nor using a simple OLS regression of hate crimes on tweets.

Figure 2.8: Trump's Tweets About Muslims and Anti-Muslim Hate Crime

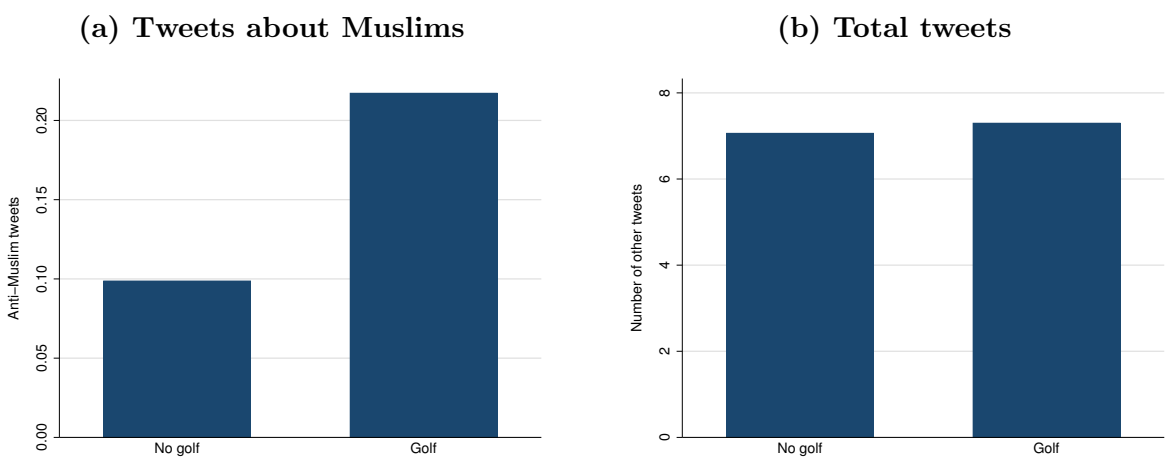


Notes: This figure plots the daily time series of anti-Muslim hate crime and Trump's tweets about Muslims, smoothed using a 14-day moving average. The time period covers the start of Trump's presidential campaign in June 2015 until the end of 2017.

We propose an instrumental variable strategy to get around the most obvious reverse causality concerns. In particular, we leverage Trump's passion for golf: in 2017 alone, Trump likely golfed on 92 days. As it turns out, the data suggest a strong link between Trump's golf outings and his Twitter feed: Figure 2.9 shows that while the total number of tweets he sends are unchanged on golf days, the *content* of his tweets sharply tilts towards negative,

Muslim-related rhetoric. In 2017, 15 out of the 34 tweets we classify as negatively mentioning Muslims were sent on golf days. In Figure 2.25 in the online appendix, we show that the topic shift is explained by a drop in policy-related tweets and more frequent mentions of Muslims and the media. Figure 2.26 shows that his tweets also become more negative in sentiment. One intuitive explanation for this pattern is that once Trump is away from the White House, his attention shifts away from policy issues. Another influence on Trump's social media activity that is likely stronger on golf days is his social media manager Dan Scavino, who is know to have suggested tweets and topics to Trump (Edwards, 2018). Figure 2.27 in the online appendix provides additional evidence that Trump's daily schedule influences the content of his tweets. In particular, we show that Trump is more likely to tweet about foreign politics when he is abroad and more likely to tweet about domestic and party politics on days he receives a policy briefing.

Figure 2.9: Trump's Twitter Activity, Split by Golf Days



Notes: These figures plot the average daily number of Trump's tweets, split by whether he plays golf on a given day in 2017. Panel (a) reports the average number of tweets about Muslims, panel (b) reports the total number of tweets.

Because the President's schedule is to some extent predetermined to accommodate security concerns and meetings, it is plausibly exogenous with respect to hate crimes against Muslims. What matters for our identification strategy is that Trump's golf outings are not systematically correlated with unobservable anti-Muslim sentiment. One disadvantage of this strategy is that we can only analyze 2017, for which we have both details about Trump's schedule and data on hate crimes. We also present OLS regressions for the IV sample and using the full time period since Trump joined Twitter in 2009 below.

More formally, we run time series regressions using the following framework:

$$Hate\ Crimes_{t+h} = \alpha + \beta \cdot Muslim\ Trump\ Tweets_t + \mathbf{X_t'}\gamma + \epsilon_t \tag{2.6}$$

$$Muslim\ Trump\ Tweets_t = \alpha + \delta \cdot I[Trump\ golfs]_t + \mathbf{X_t'}\psi + \xi_t \tag{2.7}$$

The dependent variable in equation (4) is the natural logarithm of US-wide hate crimes against Muslims at day $t + h$ (with one added inside). The main regressor of interest is the natural logarithm of the number of Donald Trump's Muslim tweets (again with one added inside). In the baseline specification, the vector X_t includes time trends and a full set of day-of-week as well as year-month fixed effects.

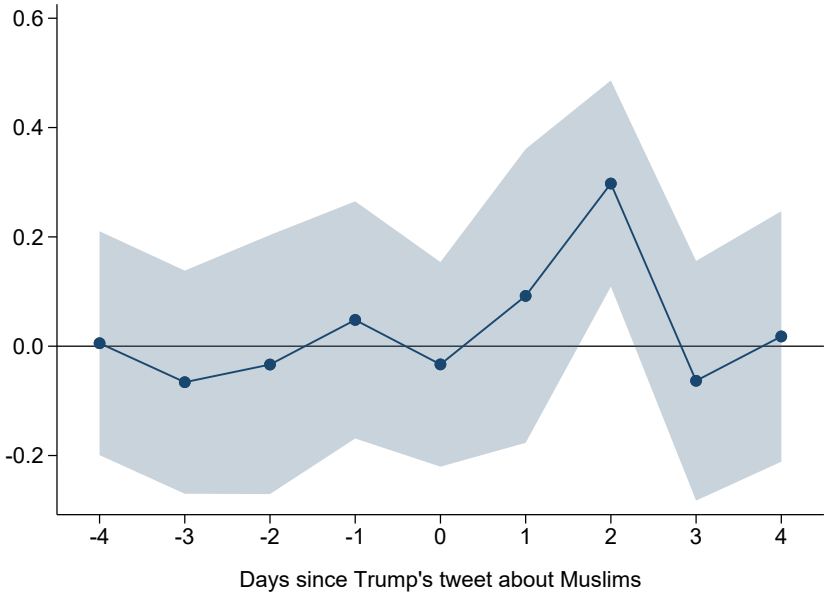
Naively estimating equation (4) would not be informative about whether Trump's Twitter activity might contribute to driving sentiments because his tweets cannot be regarded as random. We will thus instrument for tweets about Muslims in equation (5) using $I[Trump\ golfs]_t$, an indicator variable that is 1 for days on which Trump plays golf (see Section 2.2 for more details). We base inference on Newey-West standard errors that allow for heteroscedasticity and autocorrelation.

The appropriate choice of the prediction horizon h depends on the lead-lag relationship between Trump's tweets and real-life hate crimes. We plot the result from estimating equation (4) with OLS using values for h from -4 to 4 in panel (a) of Figure 2.10. As we can see, the log number of anti-Muslim hate crimes is essentially flat prior to Trump's tweets and subsequently rises to its peak in T+2. In our baseline regressions, we will thus set h to 2. We repeat the baseline estimations for different time windows in Table 2.35 in the online appendix. Panel (b) also plots the dynamic relationship between Trump's golf outings and tweets about Muslims. We can see that his tweets only increase on the days he golfs, with no similar spikes before and after.

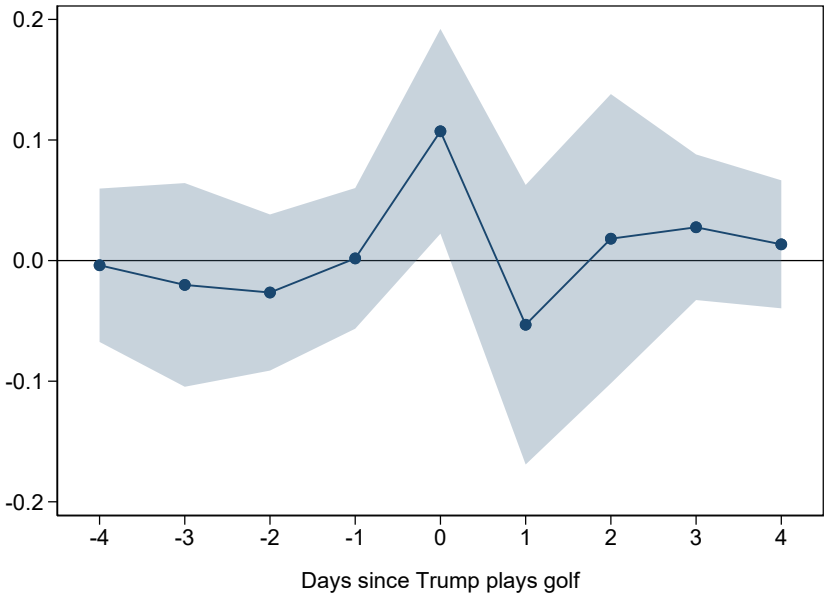
Table 2.6 presents the regression results of this exercise. We plot the OLS coefficients in panel A, first stage coefficients in panel B, reduced form coefficients in panel C, and the 2SLS estimation in panel D. Across the different specifications, the estimations suggest a clear link between Trump's tweets about Muslims and subsequent real-life hate crimes. Notably, the reduced form and 2SLS coefficients are almost fully unchanged when we include controls for measures of the salience of Muslim-related topics based on Google searches and the number of mentions on the big three TV networks (Fox News, CNN, and MSNBC). Taken at face value, this indicates that his golf outings are indeed not timed to coincide with periods of

Figure 2.10: Time Series Correlations

(a) OLS - Trump Tweets about Muslims and Hate Crimes



(b) First Stage - Golf and Trump Tweets about Muslims



Notes: These figures plot the dynamic correlations for equations 4 and 5 for values of h ranging between -4 and 4. Panel (a) plots the correlation of Trump's tweets about Islam-related topics and anti-Muslim hate crimes (both in natural logarithm). Panel (b) plots the correlation of Trump's golf outings with the log number of his Islam-related tweets. T indicates the date of tweets about Muslims or golfing ($h = 0$). All regressions include time trends; a full set of day of week and year-month dummies; and four lags of dummies for the incidence of terror attacks in the US, Europe, and the rest of the world. The sample is 2017. The shaded areas are 95% confidence intervals based on Newey-West standard errors.

high Muslim salience.

As mentioned above, a concern with instrumental variable estimation is the weak instruments problem. Because we only have one year of data to work with, this is a particular challenge in our setting. However, two pieces of information suggest that our estimates are meaningful. First, the robust F-statistics we find are consistently above the widely used linear IV rule of thumb of 10. Most of them are above the critical value for a worst case bias of 30% (at 5% statistical significance) using the cutoffs from Olea and Pflueger (2013). Second, the Anderson-Rubin confidence sets constructed using the two-step approach proposed in Andrews (2018) always exclude a zero estimate even if we assume that the instrument is weak. The reduced form and 2SLS results thus suggest that Trump's tweets could indeed be a contributing factor triggering potential perpetrators to commit real-life hate crimes.

To get a sense of the implied magnitudes, consider the estimate in column 7 of panel D Table 2.6. The coefficient of 1.659 implies that a one standard deviation increase in the log number of tweets about Muslims (0.25) is associated with a 41 log-point increase in hate crimes. This effect is large and, importantly, much larger than the OLS estimate of 0.116. An obvious explanation for this difference would be the presence of a weak instrument. However, given that the diagnostic tests discussed above are relatively encouraging, another possibility is that unobserved third factors lead to a downward bias of the OLS estimates. For example, Trump's tweets about Muslims might coincide with periods of *low* pre-existing anti-Muslim sentiment. In that case, the relationship between his tweets and hate crimes estimated via OLS would be downward biased because it conflates the true Trump effect with low general anti-Muslim sentiment. This explanation is also consistent with the finding that controlling for general attention paid to Muslims or terror attacks in columns 4 through 6 *increases* the point estimates relative to the baseline specification.

A limitation of these findings is that they are limited to the year 2017. In Table 2.38 in the online appendix, we re-run the OLS estimation for the entire period since Trump's first tweet in 2009 and split the sample into the period before and after the launch of his presidential run on June 16, 2015. We find very similar OLS estimates on his tweets about Muslims, but only after the start of his presidential campaign. For the much longer period from 2009 to mid-2015, his tweets seem to be uncorrelated with anti-Muslim hate crimes. While many factors may explain this finding, it is at least some indication that we are not capturing a phenomenon that is limited to a single year.

In Table 2.33 in the online appendix, we report more robustness results. Our results

remain largely unchanged when we control for more lags of the dependent variable to capture stronger serial correlation in hate crimes. We further experiment with additional controls for the total length of Trump's golf outings in column 3, a control if Trump golfed in the previous week (column 4), or alternative definitions of the golf dummy in columns 6 and 7. Our results are also robust to using a dummy for days with *any* Islam-related tweet from Trump (column 5).

Given the relatively short sample period, how likely would it be to find an effect if we picked golf days at random? Figure 2.24 reports the results of a randomization test for the first stage regression of Trump's tweets about Muslims on a golf dummy, where we randomly pick 92 golf days in 2017 (except the ones used in the actual variable). The distribution of the resulting t-statistics of the golf day dummy suggests that none of the placebo coefficients are close to our estimate.

We further investigate which type of anti-Muslim hate crimes drive our results. Based on the most common criteria in the FBI data, we divide anti-Muslim incidents into vandalism, theft, burglary, robbery, and assault. The results of this exercise are presented in Table 2.34 in the online appendix. Our high-frequency results appear to be mainly driven by cases of vandalism.[54]

As a simple validation exercise, we also investigate whether Trump's messages about Muslims are also correlated with hate crimes against other minorities. In particular, we consider incidents motivated by ethnicity, race, sexual orientation, or religions other than Islam. Table 2.39 plots the predictive ability of Trump's tweets about Islam-related topics for these different types of hate crimes. We only find clear-cut correlations with crimes against Muslims, not other hate crimes. This suggests that we are not merely capturing anti-minority sentiment, but rather something Muslim-specific. We also replicate this finding using simple OLS regressions for the full sample in Table 2.40. Again, we find that only hate crimes targeting Muslims are correlated with Trump's anti-Muslim tweets; the correlation with other types of hate crimes is close to zero, both before and after the start of his presidential run.

[54]Note that this does not stand in contradiction to our cross-sectional results, for which we find the largest role for assault. The daily variation we exploit here likely picks up more spontaneous anti-Muslim incidents relative to the medium-term effects in the cross-section.

2.5.2 Trump Tweets and Twitter Spillovers

We next provide evidence for the fact that Trump's negative tweets about Muslims have a direct effect on his followers. In particular, we analyze if Trump's followers become more willing to express anti-Muslim content. For this analysis we use more than 115 million tweets drawn from a random 1% sample of Trump's followers (around 630,000 users). In this dataset, we identify tweets that are retweets of Trump's negative content about Muslims, tweets that refer to Muslim-related topics but are not retweets of Trump, and tweets that contain the hashtag #BanIslam.

We continue to run time series regressions of the type in equation (4). To start, we plot dynamic correlations in Figure 2.11, where the dependent variables are different measures of tweets (in natural logarithm). The results show a clear pattern. Trump's negative tweets about Muslims are not only widely shared by his followers over the next days but also systematically followed by a spike in new content about Muslims. The time series pattern suggests no increase of anti-Muslim sentiment before Trump's tweets.

Columns 1 through 3 in Table 2.7 provide evidence that these patterns also hold when we instrument for the tweets using golf days. We focus on contemporaneous correlations, as suggested by the pattern in Figure 2.11. The reduced form and 2SLS specifications are highly statistically significant, and the weak IV confidence sets always clearly exclude zero. The 2SLS estimates suggest that a one standard deviation increase in Trump's Muslim tweets (0.25) is followed by a doubling of retweets and an almost 30% increase in new messages about Muslims that do not mention Trump. They are also followed by a 58% increase in the use of the hashtag #BanIslam by Trump followers.

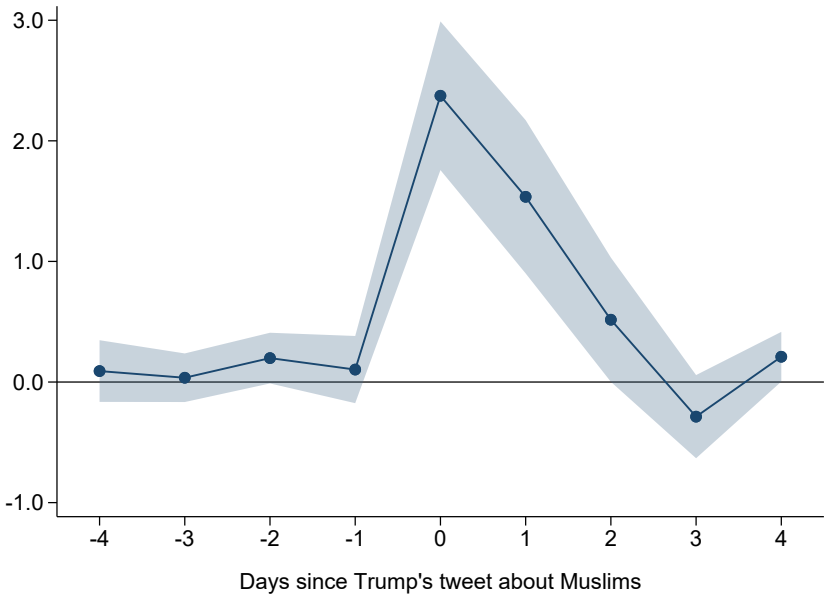
These results lend credence to the idea that Trump's tweets are trigger points for anti-Muslim sentiment among his followers. The willingness of Trump's followers to produce their own anti-Muslim messages speaks to changes in the perceived acceptability of such content after a tweet by the president.

2.5.3 Trump Tweets and the News Cycle

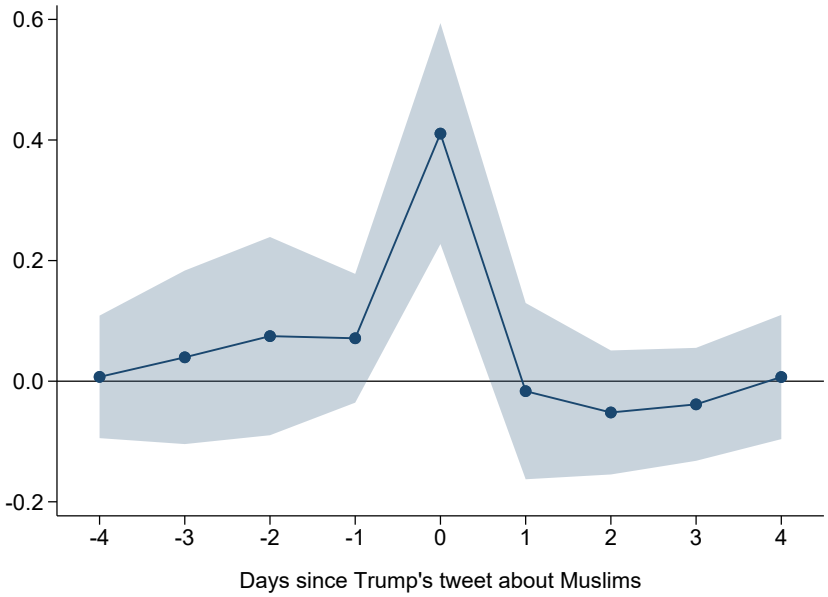
As a last time series exercise, we ask whether Trump's tweets about Muslims may have the ability to affect the news cycle. This is important to understand because, unlike for the social media channel we study here, there is ample evidence that other types of media can persuade people to participate in spontaneous, potentially violent outbursts (see e.g. DellaVigna and

Figure 2.11: Spillovers of Trump's Tweets to His Followers

(a) Retweets of Trump's Tweets



(b) New tweets about Muslims



Notes: These figures plot the dynamic correlations for equations 4 and 5 for values of h ranging between -4 and 4. Panel (a) plots the correlation of Trump's tweets about Islam-related topics and the retweets this tweets by Trump's followers (both in natural logarithm). Panel (b) plots the correlation of Trump's tweets about Islam-related topics and the self-produced anti-Muslim tweets by Trump's followers. T indicates the date of tweets about Muslims ($h = 0$). All regressions include a full set of day of week and year-month dummies; and four lags of dummies for the incidence of terror attacks in the US, Europe, and the rest of the world. The sample is 2017. The shaded areas are 95% confidence intervals based on Newey-West standard errors.

Gentzkow, 2010; Yanagizawa-Drott, 2014). As such, one obvious channel through which social media may affect offline outcomes is through influencing what other media report on. Indeed, it has been widely recognized that Twitter has become an important dissemination channel for journalists (Willnat et al., 2019); some estimates suggest that up to a quarter of Twitter users may be working for media outlets (Kamps, 2015).

We investigate the effect of Trump's tweets on media coverage using transcript data from the *TV News Archive*. In particular, we replace the dependent variable in equation (4) with the log number of mentions of Muslim-related topics on a given day by the three major cable news stations Fox News, CNN, and MSNBC. Columns 4 through 7 in Table 2.7 present the results of this exercise. Because we find a more immediate correlation between Trump's Twitter activity and news coverage, we report specifications with $h = 0$ as the prediction horizon.

Trump's tweets about Muslims are highly correlated with TV mentions in the OLS, reduced form, and 2SLS regressions. While the 2SLS estimates are still considerably larger than those obtained from OLS, they are less so than for the hate crime estimates. For overall news coverage in column 2, for example, we find that a one standard deviation increase in Muslim Trump tweets (0.25) is associated with a 74% increase in news coverage.

However, we urge caution in interpreting these results due to the short sample period. Nevertheless, the F-statistics are almost uniformly above the rule-of-thumb of 10, and mostly above the 12.04 threshold for a maximum 30% coefficient bias with 5% statistical significance derived in Olea and Pflueger (2013). Perhaps more importantly, the Anderson-Rubin confidence sets always clearly exclude zero.

We also consider heterogeneity across news stations. The correlation of instrumented Trump tweets with TV mentions appears to be strongest for Fox News (see column 5). Indeed, for CNN and MSNBC (columns 6 and 7), a zero effect is well within the AR confidence sets. This is interesting because Fox News is well-known to be supportive of Trump, following a longer term move towards more Republican-friendly reporting (Martin and Yurukoglu, 2017). This might allow Trump's comments to be broadcast uncritically and even more widely through the channel's considerable reach. Taken together, this suggests that social media may affect the news cycle, which could be one potential trigger point for potential perpetrators of hate crimes.

2.6 Panel Evidence: Trump's Tweets and Twitter Usage

As the last part of our analysis, we combine the cross sectional and time series evidence. If Trump's anti-Muslim rhetoric spreads through Twitter, we should observe large increases in anti-Muslim hate crime in counties with higher Twitter usage. We investigate this hypothesis with the following regression specification:

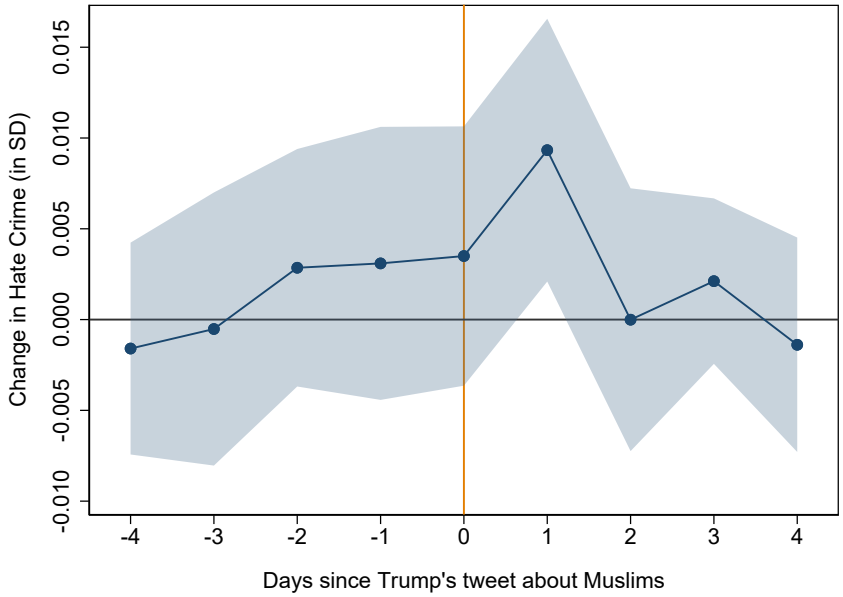
$$\begin{aligned} Hate\ Crimes_{cd} = \beta \cdot Twitter\ Usage_c \times Muslim\ Trump\ Tweets_d \\ + \mathbf{X'_{cd}}\gamma + County\ FE + Day\ FE + \epsilon_{cd} \end{aligned} \tag{2.8}$$

where $Hate\ Crimes_{cd}$ is the natural logarithm of one plus the number of hate crimes in county c on day d. The main coefficient of interest β is the interaction of county-level Twitter usage with Trump's tweets about Muslims. The coefficient measures if there are disproportionate changes in anti-Muslim hate crimes in counties with high Twitter usage on days Trump tweets about Muslims. To simplify the interpretation of the coefficients, we standardize all variables to have a mean of zero and standard deviation of one. The specification additionally controls for a vector of control variables $\mathbf{X_{cd}}$ and includes a full set of county and day fixed effects. We also allow for models that include lags of the dependent variable.[55] We cluster standard errors at the state level.

The setup in equation 2.8 is akin in spirit to a shift-share design, where $Twitter\ Usage$ measures the local exposure to aggregate shocks $Muslim\ Trump\ Tweets$. Because we are interested in estimating the effect of social media, the main concern with this empirical strategy is that the local exposure measure is co-determined with latent factors that may also lead to changes in outcomes when Trump tweets (Goldsmith-Pinkham et al., 2020). Apart from estimating equation 2.8 using OLS, we thus also present results based on 2SLS, where we again instrument for local Twitter usage using temporal fluctuations in when users started following SXSW around the 2007 festival. The exclusion restriction in this setting is that Trump's tweets about Muslims only affect areas with SXSW followers who joined in March 2007, compared to those who joined before, through its impact on Twitter usage. In support

[55] Estimates of dynamic panel models with fixed effects have an asymptotic bias of order $1/T$ (Nickell, 1981). Because we have a large T (930 days), this bias is likely negligible. Estimating the model with the GMM estimator of Arellano and Bond (1991) is not feasible because the number of moment conditions is of order T^2.

Figure 2.12: Panel Event Study – Trump Tweets, Twitter Usage, and Hate Crimes



Notes: These figures plot the dynamic correlations for equation 2.8 time periods ranging between -4 and 4 days around Trump's tweets in counties with high Twitter usage. The dependent variable is the log number of anti-Muslim hate crimes in county c on day d, which we standardized to have a mean of zero and standard deviation of one. T indicates the date of tweets about Muslims ($h = 0$). All regressions include population controls and county times month, day and county times day of month fixed effects. The shaded areas are 95% confidence intervals based on standard errors clustered at the state level.

of this, we find that the interaction of Trump's tweets with SXSW followers who joined prior to March does not predict hate crimes.[56]

We first investigate the timing of Trump's tweets with real outcomes in this panel setting. To do so, we include interactions of local Twitter usage with leads and lags of Trump's tweets about Muslims. Figure 2.12 presents the estimates of this exercise. The graph indicates that we observe differential increases in anti-Muslim hate crime in counties with high Twitter usage one day after Donald Trump's tweets. This is similar to the one we observe in the time series regression. In the online appendix in Table 2.41 we report the full set of estimated coefficients from this regressions in OLS and in reduced form.

Next, we test whether this finding is robust to the inclusion of additional fixed effects and compare the importance of Twitter usage relative to other cross-sectional predictors. In

[56]Note that these regressions are highly demanding because hate crimes are relatively rare. In these specifications, less than 1,000 of the close to three million observations are non-zero. The results should thus be interpreted with caution. Nevertheless, we believe they are insightful because they provide an additional plausibility check for the evidence presented above.

particular we analyze if exposure to Fox News or ideological alignment with Trump (measured by a high Republican vote share) are additional mediating factors.[57]

The results of these exercises can be found in Table 2.8. Overall the findings are remarkable robust to including interactions with these other cross-sectional exposure variables. The magnitude of the coefficients remains quantitatively unchanged, even when we include state × day, county × day of week and county × day of month fixed effects in columns 1-3. In the following two columns we show that the inclusion of Fox News exposure and the Republican vote share – both of which we interact with Trump's tweets – have less robust and quantitatively smaller predictive power for increases in anti-Muslim hate crime.

Overall the findings in this section are again in line with the hypothesis that, when triggered by a shock such as Trump's tweets about Muslims, social media may contribute to anti-Muslim incidents in real-life.

2.7 Discussion

2.7.1 Potential Mechanisms

The evidence provided in the previous sections all support the hypothesis that social media began to play a role in the of the expression of anti-Muslim sentiment and the spread of anti-Muslim hate crimes with the 2016 presidential campaign. The existing literature suggests that our findings could be driven by coordination, persuasion or social norms. While all mechanism are likely at play to some extent in our setting, some findings are more consistent with a role for social norms.

To begin, our findings are unlikely to be driven by lower coordination costs due to social media. The main reason is that neither the 2016 presidential campaign period nor Trump's tweets sharply improved the coordination capabilities of perpetrators of anti-Muslim hate crimes. Further, because most content on Twitter is entirely public, one would not expect it to be the most likely place for plotting anti-Muslim attacks but rather a place to spread ideas.

Another hypothesis is that our findings are driven by the persuasiveness of Twitter content, and Trump's tweets in particular (see DellaVigna and Gentzkow, 2010, for a review of the literature on persuasion). The short-lived spikes in anti-Muslim hate crime we are

[57]Note that we focus on additional cross-sectional exposure variables because we are interested in the effect of social media per se. As we show above, measures of anti-Muslim sentiment (e.g. Fox News reports) are at least partially *outcomes* of Trump's tweets.

observing in the time series are perhaps most in line with a persuasion story. But while persuasion can explain some of our findings, there are some pieces of evidence that are not easily rationalized in a belief-based persuasion model. First, in most persuasion models, the updating of beliefs depends on the credibility of the receiver as well as the information provided (Kamenica and Gentzkow, 2011). However, Trump's tweets for the most part do not contain hard information. This makes it less likely that people are persuaded to commit hate crimes against Muslims compared to the possibility that Trump's tweets trigger people with existing anti-Muslim biases. Second, belief-based models of persuasion would suggest that people with weaker priors adjust their attitudes more strongly. In contrast, we find that the effects of Twitter usage are driven by areas with *higher* pre-existing prejudice. This is also in line with existing evidence of media persuasion: in the case of Nazi radio propaganda, Adena et al. (2015) show that it predominantly activated existing sentiments (also see Voigtlander and Voth, 2012). Third, most persuasion models would predict increases in *average* anti-Muslim hostility. Panel survey evidence in Hopkins and Washington (2019), however, suggests that white Americans' anti-minority prejudice, if anything, declined after Trump's political rise.

We also provide some additional evidence that is difficult to square with the idea that social media affects violence by making people more xenophobic, at least in our setting. Table 2.30 reports the results from regressions of the type in 2.4, where the dependent variable is now the change in a measure of implicit bias against Muslims around Trump's presidential campaign start. This measure is based on mean scores on implicit association tests (IAT) from Project Implicit, which are based on the difference in an individual's ability to assign positive or negative words to Muslims or other people.[58]

We consider a range of specifications and sub-samples, including test scores restricted to whites or conservative, and find no evidence of an increase in implicit bias. In fact, both the time series mean and the estimates based on SXSW suggest that, if anything, people became *less* biased towards Muslims between 2000 and 2017. The estimates suggest that we can reject even small increases in implicit bias due to social media. The weak IV confidence set for the baseline estimate in column 1 is bounded at 0.03, which suggests we can likely rule out that a one standard deviation increase in Twitter usage increases implicit bias by more

[58]We follow Chetty et al. (2018) and calculate mean IAT scores on the county-level. Participation in the IAT is online and largely voluntary, which may give rise to selection biases. While we cannot fully rule out such biases, we also consider a measure of implicit bias based on individuals who were obligated to take these tests, e.g. as part of a work program, and find similar results.

than 17% of a standard deviation.[59] This conclusion is also supported by the pattern of the event study in Figure 2.22.

A perceived shift in social norms among people who already harbor extreme viewpoints may be an alternative mechanism to explain why we observe an effect of social media on hate crime and expressed xenophobia, but no effect on implicit biases. The channel we have in mind is the following. A key feature of social norms is that they are based on people's *perceptions* of everyone else's beliefs. These perceptions, in turn, are shaped by the "sample" of beliefs that are most salient to an individual (e.g. Bursztyn and Jensen, 2015; Perez-Truglia and Cruces, 2017; Enikolopov et al., 2020). But the people are systematically wrong in their perception of what others believe, particularly when it comes to political topics (e.g. Westfall et al., 2015; Bordalo et al., 2016).[60]

By enabling relatively few but particularly visible individuals to affect the aggregate discourse, social media could shift beliefs about what is socially acceptable and make people more susceptible to extreme viewpoints. Such effects could be re-enforced by what has often been called "echo chambers" (e.g. Bessi et al., 2015; Del Vicario et al., 2016; Schmidt et al., 2017; Sunstein, 2017). This, in turn, could affect the willingness of a small set of potential perpetrators to take hateful actions online or offline.[61]

This interpretation is in line with the findings of Bursztyn et al. (c), who show in a range of experiments that Donald Trump's 2016 election victory increased people's willingness to publicly express xenophobic views, as well as the tolerance towards such views. While our setting does not allow for a controlled experiment, our findings suggest that social media could contribute to such an unraveling of social norms.[62]

[59] To see this, consider that the standard deviation of *Log(Twitter usage)* in this sample is around 1.80. The standard deviation of the change in IAT scores is 0.313. That means the largest effect of a one standard deviation increase in social media usage in the confidence set is $(0.03 \times 1.80)/0.313 \approx 0.17$. In other words, 1% higher social media usage is unlikely to increase implicit bias against Muslims by more than 0.17%.

[60] See Bénabou (2008) for a model of how belief distortions can give rise to a persistence of ideologies in equilibrium; Bénabou (2013) studies "groupthink" more broadly. False beliefs can also result in an aggregate misperception, termed "pluralistic ignorance" (see Miller and Prentice, 1994; Kuran, 1995). In Saudi Arabia, for example, most men privately approve of women in the labor force but drastically underestimate approval among their peers (Bursztyn et al., d).

[61] This is related to Ali and Bénabou, where the visibility of individuals makes aggregate behavior (*descriptive* norms) less informative about societal preferences (*prescriptive* norms). It is also related to Mukand and Rodrik, where "political entrepreneurs" can change individuals' perception of whom they are, by increasing the salience of particular parts of their identity (e.g. a "true American"). Matz et al. (2017) provide evidence for the effectiveness of social media targeting based on psychological traits.

[62] For theoretical models of social norms see, for example, Bénabou and Tirole (2006), Bénabou and Tirole, Ali and Lin (2013), and Ali and Bénabou. Daughety and Reinganum (2010) study how agents adjust their actions if they are observable by others, which creates a costly social distortion. For empirical evidence on

2.7.2 Reporting Changes in Hate Crimes

A potential concern for interpreting our findings with regard to hate crimes could be reporting bias in the FBI data. We believe it is highly unlikely that our findings are solely driven by changes in the reporting rather than actual incidents of hate crimes.

First, our cross-sectional empirical strategy makes the most obvious types of reporting changes unlikely. We focus on within-county changes of hate crime after taking out state-level averages. This rules out any persistent differences in the propensity to report hate crimes, as well as dynamic changes across states. In our instrumental variable estimation, we exploit variation in the locations of SXSW followers who joined in March 2007, compared to those of SXSW followers from previous months. It is not clear why changes in reporting, without changes in actual hate crime incidents, would exhibit this particular correlation with early Twitter adoption. To the best of our knowledge, social media activity is not a major input in the two-tier process for the identification of hate crimes by the FBI.

Second, the heterogeneous patterns we find in the data are inconsistent with those one would expect for changes in hate crime reporting. The cross-sectional results are entirely driven by one crime category, assault. If social media only increased reporting, we would expect to see more reports on hate crimes of lower significance, such as minor cases of vandalism, which is not the case in the data. Reporting also does not explain why there should be larger effects in counties with pre-existing hate groups. If anything, one would expect reporting changes with the start of Trump's presidential run to be concentrated in more liberal counties. Further, Hobbs and Lajevardi (2019) find that the 2016 presidential election was associated with a partial withdrawal of Muslims from public life. In that case, changes in reporting would further bias our estimates downwards.

Third, the precise timing in our time series results speaks against reporting changes. While people might report more hate crimes after Trump's negative tweets about Muslims, they should also become more likely to report *past* hate crimes. This would lead to a very different time series pattern: increases in reporting should should translate into a larger number of hate crimes not only after but also *before* Trump's tweets. However, the data only shows a spike *after* the tweets. It also seems unlikely that the time series findings are driven by changes in the way the FBI classifies hate crimes, because the incident date rarely corresponds to the date a hate crime is reviewed by the FBI as part of the two-tier process.

persuasion and social norms, see e.g. Cialdini et al. (2006), Gerber et al. (2008), DellaVigna and Gentzkow (2010), and DellaVigna et al. (2016).

If Trump's tweets change the behavior of FBI analysts, this would again lead to increases in hate crimes before Trump's tweets, which we do not observe in the data.

Taken together, we believe our evidence to be more in line with changes in the actual number of hate crimes. This is also consistent with evidence using the alternative data from the Anti-Defamation League we use in robustness exercises.

2.8 Conclusion

Social media has recently come under scrutiny for its oft-alleged potential to increase citizen polarization by creating informational "echo chambers" (Sunstein, 2009, 2017). Yet, the empirical evidence on this question is limited and has led to widely differing conclusions (Boxell et al., 2017). Consistent with evidence that social media can motivate real-life action (Enikolopov et al., 2020; Müller and Schwarz, 2018a), we find a tight link between Twitter usage, Donald Trump's tweets about Muslims, and different measures of anti-minority sentiment.

Using an instrumental variable strategy, we attempt to identify the causal effect of social media on anti-Muslim sentiment around the time that then-candidate Trump launched his campaign. We exploit the unique history of the diffusion of Twitter prompted by the service's surge in popularity at the SXSW conference in March 2007. This fact allows us to instrument for social media usage today using the locations of Twitter's early adopters while holding constant the locations of people following SXSW prior to the 2007 event or other events similar to SXSW. By identifying the effect through the time dimension, this approach allows us to abstract from endogenous selection into Twitter penetration under relatively mild identifying assumptions.

Our findings are consistent with a role for social media in the normalization of anti-minority sentiments. In line with this hypothesis, we find that Trump's tweets about Muslims are highly correlated with the number of anti-Muslim hate crimes, but only for the time period after the start of his presidential campaign. This correlation also persists using an instrumental variable strategy that leverages the fact that Trump tweets more about Muslims on days when he golfs. This is at least suggestive of the idea that social media, and Trump's tweets in particular, may contribute to a climate that reduces the social sanctions against and increases the incidence of hate crimes.

While this paper focused on particularly negative outcomes – hate crimes targeting minorities and other measures that broadly reflect xenophobia – social media may well have a positive impact in other areas. We would also like to caution against using our findings as a basis for policies directed at restricting online communication. History is ripe with cautionary tales of how excessive state power over the media can abet or enable authoritarian rule. The complex trade-offs that policy makers face in this regard thus require nuanced discussion and, above all, more evidence. Notwithstanding, our results suggest that social media can affect offline actions that might endanger minority communities, and should be taken seriously.

Table 2.1: First Stage - South by Southwest 2007 and the Diffusion of Twitter Usage

	Log(Twitter usage)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log(SXSW followers, March 2007)	0.505***	0.461***	0.440***	0.407***	0.403***	0.394***	0.371***	0.362***
	(0.061)	(0.061)	(0.064)	(0.054)	(0.052)	(0.053)	(0.056)	(0.056)
Log(SXSW followers, Pre)	0.153*	0.162*	0.120	0.112	0.104	0.102	0.090	0.086
	(0.077)	(0.091)	(0.089)	(0.084)	(0.083)	(0.081)	(0.081)	(0.077)
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Population controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Race and religion controls				Yes	Yes	Yes	Yes	Yes
Socioeconomic controls					Yes	Yes	Yes	Yes
Media controls						Yes	Yes	Yes
Election control							Yes	Yes
Crime controls								Yes
Geographical controls			Yes	Yes	Yes	Yes	Yes	Yes
Observations	3107	3107	3107	3107	3106	3105	3105	3105
Mean of DV	10	10	10	10	10	10	10	10
p-value: March 2007 = Pre	0.01	0.04	0.03	0.02	0.01	0.01	0.02	0.02

Notes: This table presents county-level regressions where the dependent variable is the number of tweets sent (in natural logarithm). *SXSW followers, March 2007* is the number of Twitter users who joined in March 2007 and follow South by Southwest (SXSW) *SXSW followers, Pre* is the number of SXSW followers who registered at some point in 2006. The bottom row reports p-values from F-tests for the equality of these coefficients. All regressions control for population deciles and state fixed effects (not shown). Demographic controls include population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. Race and religion controls contains the share of people identifying as white, African American, Native American or Pacific Islander, Asian, Hispanic, or Muslim. Socioeconomic controls include the poverty rate, unemployment rate, local GINI index, the share of uninsured individuals, log median household income, the share of highschool graduates, the share of people with a graduate degree, as well as the employment shares in agriculture, information technology, manufacturing, nontradables, construction and real estate, utilities, business services, or other sectors. Media controls include the viewership share of Fox News, the cable TV spending to population ratio, and the prime time TV viewership to population ratio. Election control is the county-level vote share of the Republican party in 2012. Crime controls are the rates of violent or property crime from the FBI. Geographical controls include the linear distance from the SXSW festival location (Austin, Texas), population density, and the natural logarithm of county size. Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.2: Reduced Form - South by Southwest 2007 and the Rise in Hate Crimes against Muslims

	ΔLog(Hate crimes against Muslims)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log(SXSW followers, March 2007)	0.075**	0.074**	0.082***	0.075**	0.072**	0.072**	0.072**	0.072**
	(0.030)	(0.030)	(0.029)	(0.031)	(0.030)	(0.030)	(0.030)	(0.030)
Log(SXSW followers, Pre)	0.033	0.034	0.050	0.025	0.025	0.026	0.026	0.027
	(0.054)	(0.054)	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Population controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Race and religion controls				Yes	Yes	Yes	Yes	Yes
Socioeconomic controls					Yes	Yes	Yes	Yes
Media controls						Yes	Yes	Yes
Election control							Yes	Yes
Crime controls								Yes
Geographical controls			Yes	Yes	Yes	Yes	Yes	Yes
Observations	3107	3107	3107	3107	3106	3105	3105	3105
Mean of DV	.019	.019	.019	.019	.019	.019	.019	.019

Notes: This table presents county-level regressions where the dependent variable is the log change in hate crimes against Muslims between 2010 and 2017. *SXSW tweets* are the number of newly registered users in the indicated months of 2007 that tweeted about the South by Southwest (SXSW) festival. All regressions control for population deciles and state fixed effects (not shown). Demographic controls include population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. Race and religion controls contains the share of people identifying as white, African American, Native American or Pacific Islander, Asian, Hispanic, or Muslim. Socioeconomic controls include the poverty rate, unemployment rate, local GINI index, the share of uninsured individuals, log median household income, the share of highschool graduates, the share of people with a graduate degree, as well as the employment shares in agriculture, information technology, manufacturing, nontradables, construction and real estate, utilities, business services, or other sectors. Media controls include the viewership share of Fox News, the cable TV spending to population ratio, and the prime time TV viewership to population ratio. Election control is the county-level vote share of the Republican party in 2012. Crime controls are the rates of violent or property crime from the FBI. Geographical controls include the linear distance from the SXSW festival location (Austin, Texas), population density, and the natural logarithm of county size. Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.3: 2SLS - Social Media and the Rise in Hate Crimes against Muslims

	ΔLog(Hate crimes against Muslims)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: OLS - Hate crimes against Muslims								
Log(Twitter usage)	0.021***	0.019***	0.019***	0.015***	0.015***	0.016***	0.015***	0.015***
	(0.006)	(0.006)	(0.007)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)
Panel B: 2SLS - Hate crimes against Muslims								
Log(Twitter usage)	0.148**	0.161**	0.187**	0.185**	0.178**	0.183**	0.194**	0.199**
	(0.064)	(0.069)	(0.075)	(0.082)	(0.080)	(0.083)	(0.091)	(0.093)
Weak IV 95% AR confidence set	[0.04; 0.27]	[0.04; 0.30]	[0.06; 0.35]	[0.04; 0.35]	[0.04; 0.34]	[0.04; 0.35]	[0.04; 0.39]	[0.04; 0.40]
Log(SXSW followers, Pre)	0.010	0.008	0.027	0.005	0.007	0.008	0.009	0.010
	(0.065)	(0.069)	(0.065)	(0.064)	(0.062)	(0.062)	(0.062)	(0.061)
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Population controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Race and religion controls				Yes	Yes	Yes	Yes	Yes
Socioeconomic controls					Yes	Yes	Yes	Yes
Media controls						Yes	Yes	Yes
Election control							Yes	Yes
Crime controls								Yes
Geographical controls			Yes	Yes	Yes	Yes	Yes	Yes
Observations	3107	3107	3107	3107	3106	3105	3105	3105
Mean of DV	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019
Robust F-stat.	68.03	58.04	46.96	56.25	61.27	55.30	43.89	41.82

Notes: This table presents county-level OLS and IV regressions where the dependent variable is the log change in hate crimes against Muslims between 2010 and 2017. *Log(Twitter usage)* is instrumented using the number of users who started following SXSW in March 2007. *SXSW followers, Pre* is the number of SXSW followers who registered at some point in 2006. All regressions control for population deciles and state fixed effects (not shown). Demographic controls include population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. Race and religion controls contains the share of people identifying as white, African American, Native American or Pacific Islander, Asian, Hispanic, or Muslim. Socioeconomic controls include the poverty rate, unemployment rate, local GINI index, the share of uninsured individuals, log median household income, the share of highschool graduates, the share of people with a graduate degree, as well as the employment shares in agriculture, information technology, manufacturing, nontradables, construction and real estate, utilities, business services, or other sectors. Media controls include the viewership share of Fox News, the cable TV spending to population ratio, and the prime time TV viewership to population ratio. Election control is the county-level vote share of the Republican party in 2012. Crime controls are the rates of violent or property crime from the FBI. Geographical controls include the linear distance from the SXSW festival location (Austin, Texas), population density, and the natural logarithm of county size. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) using the Stata package from Sun (2018). For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.4: Further Robustness - Social Media and the Rise in Hate Crimes against Muslims

	No state FE (1)	Change since 1990 (2)	Log hate crimes (3)	Drop zero change counties (4)	Drop potentially nonreporting counties (5)	Drop counties with few Muslims (6)	Population weights (7)	Only neighbouring counties (8)	Index dependent variable (9)
Panel A: OLS - Hate crimes against Muslims									
Log(Twitter usage)	0.012*	0.028***	0.067***	0.042	0.028***	0.057**	0.087**	0.044***	0.031**
	(0.006)	(0.008)	(0.017)	(0.032)	(0.009)	(0.024)	(0.040)	(0.013)	(0.015)
Panel B: Reduced form - Hate crimes against Muslims									
Log(SXSW followers, March 2007)	0.072**	0.125***	0.224***	0.112**	0.071**	0.080**	0.113***	0.079**	0.172**
	(0.030)	(0.031)	(0.045)	(0.047)	(0.032)	(0.034)	(0.038)	(0.033)	(0.072)
Panel C: 2SLS - Hate crimes against Muslims									
Log(Twitter usage)	0.154**	0.271***	0.487***	0.234**	0.142**	0.173**	0.210***	0.181**	0.373**
	(0.065)	(0.069)	(0.104)	(0.103)	(0.065)	(0.075)	(0.068)	(0.084)	(0.169)
Weak IV 95% AR confidence set	[0.03; 0.28]	[0.15; 0.41]	[0.30; 0.69]	[0.06; 0.43]	[0.02; 0.27]	[0.03; 0.32]	[0.08; 0.34]	[0.04; 0.36]	[0.09; 0.71]
Log(SXSW followers, Pre)	0.019	-0.021	0.051	0.019	0.032	0.041	-0.020	0.010	-0.066
	(0.066)	(0.071)	(0.117)	(0.089)	(0.064)	(0.070)	(0.058)	(0.074)	(0.156)
Observations	3108	3107	3107	381	2319	586	3107	1167	3107
Mean of DV	0.019	0.025	0.052	0.153	0.026	0.082	0.155	0.040	0.029
Robust F-stat.	80.40	58.04	58.04	64.79	80.13	61.35	44.91	37.76	58.04

Notes: This table presents county-level OLS and IV regressions where the dependent variable is the log change in hate crimes against Muslims between 2010 and 2017 in all columns except columns 2 and 3. In column 2, the dependent variable is the log change between 1990 and 2017; in column 3, it is the log number of hate crimes against Muslims in a county after the start of Donald Trump's presidential run on June 16, 2015. *Log(Twitter usage)* is instrumented using the number of users who started following SXSW in March 2007. *SXSW followers, Pre* is the number of SXSW followers who registered at some point in 2006. All regressions control for population deciles, state fixed effects (except in column 1), and demographic controls that include population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. Column 4 drops all counties for which the change in hate crimes between 2010 and 2017 was zero. Column 5 drops all counties which never report a hate crime between 1990 and 2017. Column 6 drops all counties for which the (rounded) share of Muslims in the county population is zero according to Census data. Column 7 estimates all regressions using weighted least squares (WLS) with population weights. Column 8 only keeps neighbouring counties that differ in whether they have SXSW followers in March 2007 or not. Column 9 recodes the dependent variable into an index equal to 1 for increases in hate crimes, -1 for decreases, and 0 for no change. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) using the Stata package from Sun (2018). For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.5: Social Media and Other Hate Crimes

	FBI Data						ADL Data
	Total (1)	Hispanic (2)	Other ethnic (3)	Race (4)	Sexual Orientation (5)	Religion (excl. Muslims) (6)	Total (Levels) (7)
Panel A: OLS - Hate crimes							
Log(Twitter usage)	0.006	-0.000	-0.018***	0.005	-0.007	0.017*	0.129***
	(0.012)	(0.008)	(0.007)	(0.008)	(0.006)	(0.009)	(0.034)
Panel B: Reduced form - Hate crimes							
Log(SXSW followers, March 2007)	0.085**	0.079**	0.007	0.055	0.046	0.058	0.357***
	(0.042)	(0.034)	(0.033)	(0.048)	(0.043)	(0.041)	(0.110)
Panel C: 2SLS - Hate crimes							
Log(Twitter usage)	0.184*	0.171**	0.014	0.119	0.099	0.125	0.775***
	(0.100)	(0.068)	(0.071)	(0.109)	(0.096)	(0.084)	(0.192)
Weak IV 95% AR confidence set	[0.01; 0.40]	[0 .0; 0.29]	[0.10; 0.16]	[0.06; 0.34]	[0.07; 0.29]	[0.04; 0.27]	[0.38; 1.13]
Log(SXSW followers, Pre)	-0.052	-0.074	-0.039	-0.035	-0.025	-0.036	0.055
	(0.078)	(0.071)	(0.074)	(0.081)	(0.082)	(0.064)	(0.177)
Observations	3107	3107	3107	3107	3107	3107	3107
Mean of DV	-0.015	-0.012	-0.016	-0.011	-0.025	0.005	0.226
Robust F-stat.	58.04	58.04	58.04	58.04	58.04	58.04	58.04

Notes: This table presents county-level OLS, reduced form, and IV regressions where the dependent variable is the log change in hate crimes against the group in the top row between 2010 and 2017. *Log(Twitter usage)* is instrumented using the number of users who started following SXSW in March 2007. All regressions control for population deciles and state fixed effects (not shown). Demographic controls include population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. The hate crime data from the Anti-Defamation League (ADL) is sparse prior to 2016, so we use the log-level of hate crimes in column 7. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) using the Stata package from Sun (2018). For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.6: Trump Tweets and Anti-Muslim Hate Crimes

	Baseline (1)	Add lagged dependent variable (2)	Add federal holiday control (3)	Add Google search control (4)	Add TV coverage control (5)	Add terror attack control (6)	Add total tweets control (7)
Panel A: OLS - Log(Hate crimes against Muslims) in t+2							
Log(Muslim Trump tweets)	0.130*	0.140**	0.132*	0.101	0.099	0.192**	0.116
	(0.069)	(0.066)	(0.068)	(0.062)	(0.063)	(0.077)	(0.074)
Panel B: First Stage - Log(Trump tweets about Muslims)							
Trump golfs	0.102***	0.098***	0.104***	0.103***	0.078***	0.086***	0.098***
	(0.027)	(0.026)	(0.027)	(0.027)	(0.025)	(0.025)	(0.027)
Panel C: Reduced form - Log(Hate crimes against Muslims) in t+2							
Trump golfs	0.165**	0.173**	0.158**	0.168**	0.157**	0.172**	0.163**
	(0.071)	(0.076)	(0.070)	(0.068)	(0.070)	(0.074)	(0.071)
Panel D: 2SLS - Log(Hate crimes against Muslims) in t+2							
Log(Muslim Trump tweets)	1.617**	1.756**	1.523**	1.626**	2.009*	2.011*	1.659**
	(0.779)	(0.892)	(0.736)	(0.761)	(1.198)	(1.050)	(0.842)
Weak IV 95% AR confidence set	[0.31; 4.01]	[0.43; 4.49]	[0.29; 3.64]	[0.50; 3.96]	[0.47; 6.87]	[0.48; 5.79]	[0.41; 4.41]
Fixed effects (month, day of week)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trends	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	363	363	363	363	363	363	363
R^2	0.21	0.17	0.25	0.21	0.08	0.12	0.20
Robust F-stat.	13.15	12.97	13.55	13.54	9.487	10.90	11.87

Notes: This table presents OLS and IV regressions where the dependent variable is the number of hate crimes against Muslims on any given day based on FBI data. We use a dummy for days on which President Donald Trump golfs used as an instrument for his tweets about Muslims. Column 2 controls for one lag of the dependent variable and column 3 for a dummy that tags federal holidays. Column 4 controls for the first principal component of Google searches for Islam-related terms. Column 5 controls for the number of times Fox News, CNN or MSNBC mention Islam-related words in their reporting on a given day. Column 6 controls for the number of terror attacks in the US, Europe, or other countries. Column 7 controls for the total number of tweets by Donald Trump. The sample year is 2017, for which we have information on Trump's golfing. All regressions include day-of-week and year-month dummies, linear and quadratic time trends as well as a dummy for whether Trump's golfing is the first of a series of golf days. See online appendix for more details on data and variable construction. Newey-West standard errors are reported in parentheses. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) with the Stata package from Sun (2018). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.7: Spillover Effects on Trump's Followers and Cable News Coverage

	Trump followers' Muslim tweets			Cable news coverage			
	Trump retweets (1)	New content (2)	Contains #BanIslam (3)	All stations (4)	Fox News (5)	CNN (6)	MSNBC (7)
Panel A: OLS - Log(Total number of Muslim TV mentions/tweets)							
Log(Muslim Trump tweets)	2.658***	0.680***	0.360***	0.677***	0.607***	0.808***	0.660***
	(0.346)	(0.105)	(0.094)	(0.089)	(0.117)	(0.109)	(0.084)
Panel B: Reduced Form - Log(Total number of Muslim TV mentions/tweets)							
Trump golfs	0.456**	0.117**	0.234***	0.273**	0.296**	0.285	0.185*
	(0.208)	(0.058)	(0.074)	(0.134)	(0.115)	(0.212)	(0.110)
Panel C: 2SLS - Log(Total number of Muslim TV mentions/tweets)							
Log(Muslim Trump tweets)	4.508***	1.151**	2.313**	2.701**	2.923***	2.813	1.830**
	(1.305)	(0.469)	(0.955)	(1.114)	(0.966)	(1.891)	(0.921)
Weak IV 95% AR confidence set	[1.01; 6.96]	[0.17; 2.21]	[0.89; 5.43]	[0.39; 5.24]	[1.11; 5.31]	[-1.49; 7.12]	[-0.27; 3.93]
Fixed effects (month, day of week)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trends	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	364	364	364	364	364	364	364
Robust F-stat.	13.02	13.02	13.02	13.02	13.02	13.02	13.02

Notes: This table presents OLS and IV regressions where the dependent variable is the number of tweets by Trump followers in columns 1 to 3 and the number of times Muslims are mentioned on cable news stations on a given day in columns 4 to 7. We use a dummy for days on which President Donald Trump golfs used as an instrument for his tweets about Muslims. *Trump retweets* are retweets by Trump followers of Trump's negative tweets about Muslims. *New content* refers to tweets by Trump followers mentioning Muslims that are no Trump retweets and do not mention Trump. *Contains #BanIslam* is the number of tweets by Trump followers containing the hashtag #BanIslam. *Cable news coverage* is based on the mentions of Muslim-related words on Fox News, CNN, and MSNBC, which are also reported separately. The sample year is 2017, for which we have information on Trump's golfing. All regressions include day-of-week and year-month dummies, linear and quadratic time trends as well as a dummy for whether Trump's golfing is the first of a series of golf days. Newey-West standard errors are reported in parentheses. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) with the Stata package from Sun (2018). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.8: Robustness Bartik Interactions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel C: OLS – Log(Hate crimes against Muslims) in t+1							
Muslim Trump Tweet × Twitter Usage	0.013**	0.010**	0.014**	0.014**	0.015**	0.017***	0.015**
	(0.005)	(0.004)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Muslim Trump Tweet × Fox News Viewership						0.002**	
						(0.001)	
Muslim Trump Tweet × Republican Vote Share 2012							-0.000
							(0.001)
Panel B: Reduced Form – Log(Hate crimes against Muslims) in t+1							
Muslim Trump Tweet × SXSW Treat	0.010**	0.009**	0.010**	0.010**	0.010**	0.010**	0.010**
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Muslim Trump Tweet × SXSW Pre	0.001	-0.000	-0.000	-0.001	-0.000	-0.000	-0.000
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Muslim Trump Tweet × Fox News Viewership						0.001*	
						(0.001)	
Muslim Trump Tweet × Republican Vote Share 2012							-0.001
							(0.001)
Panel C: 2SLS – Log(Hate crimes against Muslims) in t+1							
Muslim Trump Tweet × Twitter Usage	0.143***	0.124**	0.137**	0.141**	0.147***	0.185***	0.193**
	(0.049)	(0.052)	(0.053)	(0.053)	(0.053)	(0.068)	(0.073)
Muslim Trump Tweet × SXSW Pre	-0.005	-0.006	-0.006	-0.007	-0.007	-0.008	-0.008
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)
Muslim Trump Tweet × Fox News Viewership						0.021***	
						(0.007)	
Muslim Trump Tweet × Republican Vote Share 2012							0.024**
							(0.010)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pop. deciles x Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County x Month FE		Yes	Yes	Yes	Yes	Yes	Yes
State x Day FE			Yes	Yes	Yes	Yes	Yes
County x Day of Week FE				Yes	Yes	Yes	Yes
County x Day of Month FE					Yes	Yes	Yes
Lag dep. variable	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2887332	2887332	2886403	2886403	2886403	2885474	2886403

Notes: This table presents OLS, reduced form and IV regressions where the dependent variable is the log number of anti-Muslims hate crime in county c on day d. The independent variable is either the interaction Trump's anti-Muslim tweet with county-level Twitter usage or a reduced form/2SLS specification with our SXSW variables. The variables are standardized to have a mean of zero and standard deviation of one. All regressions include population controls, one lag of the dependent variable, as well as county and day fixed effects. Some regressions further control for county × month, state × day, county × day-of-week, and county × day-of-month fixed effects (as indicated). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

2.9 Online Appendix:

2.9.1 Appendix 1: Additional Details on Data

Table 2.9: Variable Descriptions (Part 1/2)

Variable	Description	Source
	Hate crime variables	
Hate crimes	Total number of hate crimes recorded in the FBI hate crime data.	FBI Hate Crime Data
Anti-Muslim hate crimes	Anti-Muslim hate crimes recorded in the FBI hate crime data, based on bias motivation code 24.	FBI Hate Crime Data
Anti-Hispanic hate crimes	Anti-Hispanic hate crimes recorded in the FBI hate crime data, based on the bias motivation codes 32.	FBI Hate Crime Data
Other ethnic-based hate crimes	Anti-ethnic hate crimes recorded in the FBI hate crime data, based on the bias motivation codes 33.	FBI Hate Crime Data
Anti-racial hate crimes	Racial hate crimes recorded in the FBI hate crime data, based on bias motivation codes 11, 12, 13, 14, 15, 16.	FBI Hate Crime Data
Anti-religious hate crimes	Anti-religious hate crimes (except anti-Muslim) recorded in the FBI hate crime data, based on bias motivation codes 21, 22, 23, 25, 26, 27, 28, 29, 81, 82, 83, 84, 85.	FBI Hate Crime Data
Anti-sexual orientation hate crimes	Hate crimes based on sexual orientation recorded in the FBI hate crime data, based on the bias motivation codes 41, 42, 43, 44, 45.	FBI Hate Crime Data
	Twitter data	
Trump tweets	The total number of tweets from Donald Trump's Twitter account.	Trump Twitter Archive
Muslim tweets	The number of tweets from Donald Trump's Twitter account about Islam-related topics. We start classifying these tweets by searching for the terms "sharia", "refugee", "mosque", "muslim", "islam" and "terror". We then read all tweets and verify that they indeed mention Muslims in a negative way.	Trump Twitter Archive
Twitter usage	The number of geolocated tweets per county that were collected using the Twitter streaming API in a 12 month period from June to November 2014 and June to November 2015.	Gesis Datatorium
SXSW followers, March 2007	The number of Twitter users following the SXSW account in each county that signed up to Twitter in March 2007.	Twitter Search API
SXSW followers, Pre	The total number of Twitter users following the SXSW account in each county that signed up to Twitter at any point in 2006.	Twitter Search API
Burning Man Twitter Users, August 2007	The number of Twitter users in each county that tweeted about the Burning Man festival in August 2007 and joined Twitter in August 2007.	Twitter Search API
Coachella Twitter Users, April 2007	The number of Twitter users in each county that tweeted about the Coachella festival in April 2007 and joined Twitter in April 2007.	Twitter Search API
Lollapalooza Twitter Users, August 2007	The number of Twitter users in each county that tweeted about the Lollapalooza festival in August 2007 and joined Twitter in August 2007.	Twitter Search API
	Trump golf data	
Trump golfs	A dummy variable for each day in 2017 Trump spent on a golf course and likely played golf.	NYT, trumpgolfcount.com and Pres. Schedule
Trump golfs (NYT only)	A dummy variable for each day in 2017 Trump spent on a Golf course and likely golfed, based solely on the information of the New York Times.	NYT
Trump golf (alternative)	A dummy variable for each day in 2017 Trump spent on a golf course and likely golfed, based on the information of trumpgolfcount.com and extended with information from the Pres. Schedule	trumpgolfcount.com and Pres. Schedule
Golf holiday	A dummy for any of Trump's golf outings that lasts longer than 3 days.	NYT and trumpgolfcount.com
Golf at any point in previous week	A dummy variable which is 1 if Trump golfed at any point in the previous week.	NYT and trumpgolfcount.com

Table 2.10: Variable Descriptions (Part 2/2)

Variable	Description	Source
	Other cross sectional controls	
Demographic controls	Contain the share of people in the age buckets 20-24, 25-29, 30-34, 40-44, 45-49 and 50+, and the percentage change in population between 2000 and 2016.	US Census
Education controls	Contains the share of people over 25 with at least a high school degree and the share of people over 25 with at least a graduate degree.	US Census
Race and religion controls	Contains population shares of Muslims, Whites, Blacks, Native Americans, Asians, and Hispanics.	US Census/Religious Census
Socioeconomic controls	Contains a county's poverty rate, unemployment rate, GINI coefficient, share of uninsured, log of median household income, and the share of the population employed in agriculture, manufacturing, accommodation/retail, utilities, information technologies services, and other industries.	US Census/Bureau of Labor Statistics
Media controls	Contains the ratio of prime time TV viewership to population, cable spending to population, and the share of Fox News viewership.	SimplyAnalytics
Election control	Contains the vote share of the Republican party in the 2012 presidential election.	MIT Election Lab
Crime controls	Contains the number of violent crime per capita as well as the number of property crimes per capita based on FBI data.	FBI UCR Data
Distance control	Contains the distance to Austin Texas, the population density, and the logarithm of the land area for each county.	US Census Tigerline File
Change in implicit bias against Muslims	The change in the county-level mean implicit association test score from the Arab-Muslim module between 2015-2017 compared to 2010-2014.	Project Implicit
	Other time series variables	
Trump followers' retweets	The number of retweets of Trump's tweets about Muslims by his Twitter followers	Twitter
Trump followers' new content	The number of tweets by Trump followers containing the words "sharia", "refugee", "mosque", "muslim", "islam" or "terror".	Twitter
Contains #BanIslam	The number of tweets by Trump followers containing the term "#BanIslam".	Twitter
Muslim mentions (total)	The total number of cable news reports mentioning one of the following terms in their closed captions: "sharia", "refugee", "mosque", "muslim", "islam" and "terror".	Internet Archive
Muslim mentions (Fox News)	The total number of news reports on Fox News mentioning one of the following terms in their closed captions: "sharia", "refugee", "mosque", "muslim", "islam" and "terror".	Internet Archive
Muslim mentions (CNN)	The total number of news reports on CNN mentioning one of the following terms in their closed captions: "sharia", "refugee", "mosque", "muslim", "islam" and "terror".	Internet Archive
Muslim mentions (MSNBC)	The total number of news reports on MSNBC mentioning one of the following terms in their closed captions: "sharia", "refugee", "mosque", "muslim", "islam" and "terror".	Internet Archive
Google searches (PC)	The first princepal component of the rescaled Google trends for the following terms: "sharia", "refugee", "mosque", "muslim", "islam" and "terror".	Google Trends
Terror attack in the US	The number of Islamist terror attacks committed in the US.	Global Terrorism Database
Terror attack in Europe	The number of Islamist terror attacks committed in the Europe.	Global Terrorism Database
Terror attack elsewhere	The number of Islamist terror attacks committed outside of the US or Europe	Global Terrorism Database

FBI Hate Crime Data

As described in the Section 2.2, the FBI uses a two-tier decision making process for classifying hate crimes. FBI (2015) describes the decision making process in the following way:

> "Once the development of this collection was complete, the FBI UCR Program surveyed state UCR Program managers on hate crime collection procedures used at various law enforcement agencies which collected hate crime data employing a two-tier decision-making process. The first level is the law enforcement officer who initially responds to the alleged hate crime incident, i.e., the "responding officer" (or "first-level judgment officer"). It is the responsibility of the responding officer to determine whether there is any indication that the offender was motivated by bias. If a bias indicator is identified, the officer designates the incident as a "suspected bias-motivated crime" and forwards the case file to a "second-level judgment officer/unit." (In smaller agencies this is usually a person specially trained in hate crime matters, while in larger agencies it may be a special unit.) It is the task of the second-level judgment officer/unit to review the facts of the incident and make the final determination of whether a hate crime has actually occurred. If so, the incident is to be reported to the FBI UCR Program as a bias-motivated crime." (FBI, 2015, pp. 2-3)

As indicated, all decisions by the responding officer will be passed on for review to a second examiner. The FBI manual also outlines criteria that have to be full-filled for a crime to be classified as a hate crime:

> "An important distinction must be made when reporting a hate crime. The mere fact the offender is biased against the victim's actual or perceived race, religion, disability, sexual orientation, ethnicity, gender, and/or gender identity does not mean that a hate crime was involved. Rather, the offender's criminal act must have been motivated, in whole or in part, by his or her bias. Motivation is subjective, therefore, it is difficult to know with certainty whether a crime was the result of the offender's bias. For that reason, before an incident can be reported as a hate crime, sufficient objective facts must be present to lead a reasonable and prudent person to conclude that the offender's actions were motivated, in whole

or in part, by bias. While no single fact may be conclusive, facts such as the following, particularly when combined, are supportive of a finding of bias:

1. The offender and the victim were of a different race, religion, disability, sexual orientation, ethnicity, gender, and/or gender identity. For example, the victim was African American and the offender was white.
2. Bias-related oral comments, written statements, or gestures were made by the offender indicating his or her bias. For example, the offender shouted a racial epithet at the victim.
3. Bias-related drawings, markings, symbols, or graffiti were left at the crime scene. For example, a swastika was painted on the door of a synagogue, mosque, or LGBT center.
4. Certain objects, items, or things which indicate bias were used. For example, the offenders wore white sheets with hoods covering their faces or a burning cross was left in front of the victim's residence.
5. The victim is a member of a specific group that is overwhelmingly outnumbered by other residents in the neighborhood where the victim lives and the incident took place.
6. The victim was visiting a neighborhood where previous hate crimes had been committed because of race, religion, disability, sexual orientation, ethnicity, gender, or gender identity and where tensions remained high against the victim's group.
7. Several incidents occurred in the same locality, at or about the same time, and the victims were all of the same race, religion, disability, sexual orientation, ethnicity, gender, or gender identity.
8. A substantial portion of the community where the crime occurred perceived that the incident was motivated by bias.
9. The victim was engaged in activities related to his or her race, religion, disability, sexual orientation, ethnicity, gender, or gender identity. For example, the victim was a member of the National Association for the Advancement of Colored People (NAACP) or participated in an LGBT pride celebration.

10. The incident coincided with a holiday or a date of significance relating to a particular race, religion, disability, sexual orientation, ethnicity, gender, or gender identity, e.g., Martin Luther King Day, Rosh Hashanah, or the Transgender Day of Remembrance.
11. The offender was previously involved in a similar hate crime or is a hate group member.
12. There were indications that a hate group was involved. For example, a hate group claimed responsibility for the crime or was active in the neighborhood.
13. A historically-established animosity existed between the victim's and the offender's groups.
14. The victim, although not a member of the targeted racial, religious, disability, sexual orientation, ethnicity, gender, or gender identity group, was a member of an advocacy group supporting the victim group."

(FBI, 2015, pp. 6-7)

We report the full list of FBI bias motivation categories in Table 2.12. The hate crime categories we use in the paper are defined as follows:

Table 2.11: FBI Hate Crimes Codes

Hate Crime Category	**FBI Codes**
Muslim	24
Hispanic	32
Other ethnic	33
Racial	11, 12, 13, 14, 15, 16
Sexual orientation	41, 42, 43, 44, 45
Religious (excluding Muslim)	21, 22, 23, 25, 26, 27, 28, 29, 81, 82, 83, 84, 85

Table 2.12: Full List of FBI Bias Motivation Categories

Bias category	Bias motivation and code
Race/Ethnicity/Ancestry	Anti-American Indian or Alaska Native (13) Anti-Arab (31) Anti-Asian (14) Anti-Black or African American (12) Anti-Hispanic or Latino (32) Anti-Multiple Races, Group (15) Anti-Native Hawaiian or Other Pacific Islander (16) Anti-Other Race/Ethnicity/Ancestry (33) Anti-White (11)
Religion	Anti-Buddhist (83) Anti-Catholic (22) Anti-Eastern Orthodox (81) Anti-Hindu (84) Anti-Islamic (Muslim) (24) Anti-Jehovah's Witness (29) Anti-Jewish (21) Anti-Mormon (28) Anti-Multiple Religions, Group (26) Anti-Other Christian (82) Anti-Other Religion (25) Anti-Protestant (23) Anti-Sikh (85) Anti-Atheism/Agnosticism (27)
Sexual Orientation	Anti-Bisexual (45) Anti-Gay (Male) (41) Anti-Heterosexual (44) Anti-Lesbian (42) Anti-Lesbian, Gay, Bisexual, or Transgender (Mixed Group)
Disability	Anti-Mental Disability (52) Anti-Physical Disability (51)
Gender	Anti-Female (62) Anti-Male (61)
Gender Identity	Anti-Gender Nonconforming (72) Anti-Transgender (71)

Notes: This table reports the complete list of hate crime bias motivations as classified by the FBI. The table is reproduced from (FBI, 2015, p. 5).

Trump Twitter Data

Table 2.13: Examples of Trump's Negative Tweets about Muslims

Date	Text	Retweets
12/10/2015	"mimi_saulino: seanhannity @FoxNews Syrian Muslims escorted into U.S. through Mexico. Now arriving to Oklahoma and Kansas! Congress?"	1223
14/11/2015	Why won't President Obama use the term Islamic Terrorism? Isn't it now, after all of this time and so much death, about time!	6924
15/11/2015	"thewatcher23579: One of Paris terrorist came as Syrian refugee. Donald Trump is right again. BOMB THEIR OIL - TAKE AWAY THEIR FUNDING"	2165
17/11/2015	Refugees from Syria are now pouring into our great country. Who knows who they are - some could be ISIS. Is our president insane?	16285
22/11/2015	We better get tough with RADICAL ISLAMIC TERRORISTS, and get tough now, or the life and safety of our wonderful country will be in jeopardy!	5172
25/11/2015	I LIVE IN NEW JERSEY; @realDonaldTrump IS RIGHT: MUSLIMS DID CELEBRATE ON 9/11 HERE! WE SAW IT! https://t.co/1SksZU9qlj	2252
07/12/2015	Obama said in his speech that Muslims are our sports heroes. What sport is he talking about, and who? Is Obama profiling?	9600
07/12/2015	Statement on Preventing Muslim Immigration: https://t.co/HCWU16z6SR https://t.co/d1dhaIs0S7	4716
10/12/2015	The United Kingdom is trying hard to disguise their massive Muslim problem. Everybody is wise to what is happening, very sad! Be honest.	6028
10/12/2015	In Britain, more Muslims join ISIS than join the British army. https://t.co/LQVNz7b2Eb	4325
17/01/2016	Far more killed than anticipated in radical Islamic terror attack yesterday. Get tough and smart U.S., or we won't have a country anymore!	4126
27/03/2016	Another radical Islamic attack, this time in Pakistan, targeting Christian women & children. At least 67 dead,400 injured. I alone can solve	11353
22/05/2016	Crooked Hillary wants a radical 500% increase in Syrian refugees. We can't allow this. Time to get smart and protect America!	9758
12/06/2016	Appreciate the congrats for being right on radical Islamic terrorism, I don't want congrats, I want toughness & vigilance. We must be smart!	27146
13/06/2016	In my speech on protecting America I spoke about a temporary ban, which includes suspending immigration from nations tied to Islamic terror.	13026
25/06/2016	We must suspend immigration from regions linked with terrorism until a proven vetting method is in place.	11726
28/07/2016	Hillary's refusal to mention Radical Islam, as she pushes a 550% increase in refugees, is more proof that she is unfit to lead the country.	20106
18/10/2016	Thank you Colorado Springs. If I'm elected President I am going to keep Radical Islamic Terrorists out of our count... https://t.co/N74UK73RLK	12904
19/10/2016	ISIS has infiltrated countries all over Europe by posing as refugees, and @HillaryClinton will allow it to happen h... https://t.co/MmeW2qsTQh	16130
11/02/2017	Our legal system is broken! "77% of refugees allowed into U.S. since travel reprieve hail from seven suspect countries." (WT) SO DANGEROUS!	23082
17/08/2017	Study what General Pershing of the United States did to terrorists when caught. There was no more Radical Islamic Terror for 35 years!	30534
18/08/2017	Radical Islamic Terrorism must be stopped by whatever means necessary! The courts must give us back our protective rights. Have to be tough!	37669
15/09/2017	Loser terrorists must be dealt with in a much tougher manner.The internet is their main recruitment tool which we must cut off & use better!	21411
20/10/2017	Just out report: "United Kingdom crime rises 13% annually amid spread of Radical Islamic terror." Not good, we must keep America safe!	29854
01/11/2017	NYC terrorist was happy as he asked to hang ISIS flag in his hospital room. He killed 8 people, badly injured 12. SHOULD GET DEATH PENALTY!	43455

Notes: This table reports examples of Trump's negative tweets about Muslims, including the date of the tweet and the number of retweets the tweet received.

Table 2.14: Misclassified Trump's Anti-Muslim Tweets

Date	Text	Retweets
12/12/2012	Watching Pyongyang terrorize Asia today is just amazing!	77
26/03/2013	The Scottish windfarm was conceived by the same mind that released terrorist al-Megrahi for humanitarian reasons. ..	101
23/04/2013	Did the Boston terrorists register their guns? No. Another example of why gun control legislation is not the answer!	1192
22/09/2013	"@LebaneseKobe: @realDonaldTrump as a Muslim and as an American, i know for a fact that you Mr. Trump respect all people!	33
22/09/2013	"@mandem3: realDonaldTrump you hate muslims." Wrong	48
10/10/2013	Obama has called @GOP terrorists during this showdown. It's a shame he really doesn't think it because then he would meet all @GOP demands.	432
29/01/2014	Remember when "comedian" Bill Maher openly praised the disgusting terrorists who destroyed the World Trade Center-then got canned by ABC?	117
26/01/2015	"tomtumillo: What is worse, Geraldo screaming 'screw the terrorists' or Kenya feeling she's 'fabulous'? #CelebrityApprentice	56
15/08/2015	"javonniandjeno: realDonaldTrump AP nbc Donald Trump is Clint Eastwood, the perfect hero not scared of American terrorists. Vote Trump!"	1742
27/08/2015	"jp_sitles: realDonaldTrump HillaryClinton: she compared republicans to terrorist but will not call terrorists , terrorists. #OhMe"	2869
06/09/2015	"jasonusmc2017: blayne_troy @realDonaldTrump: He was right when he called Obama the 5 for 1 president. 5 terrorist for one no good traitor	1016
21/09/2015	"TheBrodyFile: On the Muslim issue: It might help @BarackObama if he actually supported Christians religious liberty rights.	1242
21/09/2015	"TheBrodyFile: On the Muslim issue: It might help @BarackObama if he didn't take five years to visit Israel"	818
21/11/2015	"WayneDupreeShow: "It's clear that Donald Trump was NOT even talking about a Muslim Database!" https://t.co/3tLDZj2WGV"	1020
31/12/2015	"SenSanders: I have a message for Donald Trump: No, we're not going to hate Latinos, we're not going to hate Muslims." I fully agree!	1250
23/03/2016	Just watched Hillary deliver a prepackaged speech on terror. She's been in office fighting terror for 20 years-and look where we are!	11115
23/03/2016	I will be the best by far in fighting terror. I'm the only one that was right from the beginning, & now Lyin' Ted & others are copying me.	7224
15/06/2016	I will be meeting with the NRA, who has endorsed me, about not allowing people on the terrorist watch list, or the no fly list, to buy guns.	13903
21/05/2017	Speech transcript at Arab Islamic American Summit https://t.co/eUWxJXJxbe nReplay https://t.co/VtmlSqciXx #RiyadhSummit #POTUSAbroad	11498
26/05/2017	Getting ready to engage G7 leaders on many issues including economic growth, terrorism, and security.	11322
27/05/2017	Big G7 meetings today. Lots of very important matters under discussion. First on the list, of course, is terrorism. #G7Taormina	9489
18/08/2017	Today, I signed the Global War on Terrorism War Memorial Act (#HR873.) The bill authorizes....cont https://t.co/c3zIkdtowc https://t.co/re6n0MS0cj	14892
07/09/2017	During my trip to Saudi Arabia, I spoke to the leaders of more than 50 Arab & Muslim nations about the need to confront our shared enemies.[...]	10156
11/11/2017	When will all the haters and fools out there realize that having a good relationship with Russia is a good thing, not a bad thing.[...]	39627

Notes: The table lists the tweets we excluded by hand from the set of negative Muslim tweets.

Geocoded Twitter Data

Table 2.15: Search Terms Used to Identify Users Tweeting about Other Festivals

Festival	Search Term
Austin City Limited Festival	Austin City Limits Festival
Burning Man	Burningman Burning Man
Coachella	Coachella
Electric Daisy Festival	EDC Las Vegas Electric Daisy Carnival
New Orleans Jazz and Heritage Festival	New Orleans Jazz and Heritage Festival Jazzfest
Lollapalooza	Lollapalooza
Pitchfork Music Festival	Pitchfork Music Festival Pitchforkfest
South by Southwest Festival	South by Southwest SXSW
West by Southwest Festival	West by Southwest WXSW

Table 2.16: Search Terms Used to Create a Proxy for Total Tweets

0	but	his	one	these	would
1	by	how	only	they	year
2	can	if	or	think	you
3	come	in	other	this	your
4	could	into	our	time	
5	day	it	out	two	
6	do	its	over	up	
7	even	just	people	us	
8	first	know	say	use	
9	for	like	see	want	
I	from	look	she	way	
about	get	make	so	we	
after	give	me	some	well	
all	go	most	take	what	
also	good	my	than	when	
any	have	new	that	which	
as	he	no	their	who	
at	he	not	them	with	
back	her	now	then	with	
because	him	on	there	work	

Notes: This table list the search terms we used to collect a proxy of all tweets sent from a given county.

Rescaling of Google trends

As described in Section 2.2, we use the weekly Google trends data to rescale the daily Google trend values. The daily Google trends data are scaled between 0-100 for each 90 day period, while the weekly Google trends data have a consistent scaling for the entire time period.

To arrive at consistent values, we use the following process. First, we create a scaling factor by dividing the weekly interest by the daily interest. We then multiply the daily interest data with the scaling factor. If the weekly interest is 100 and the daily interest is 25, the scaling factor will be 4 and values will be scaled up. On the other hand, if the weekly interest is low, for example 10, a daily interest of 25 would be scaled down. This way, the adjustment guarantees that daily interest will be on the same scale and thus comparable over time.

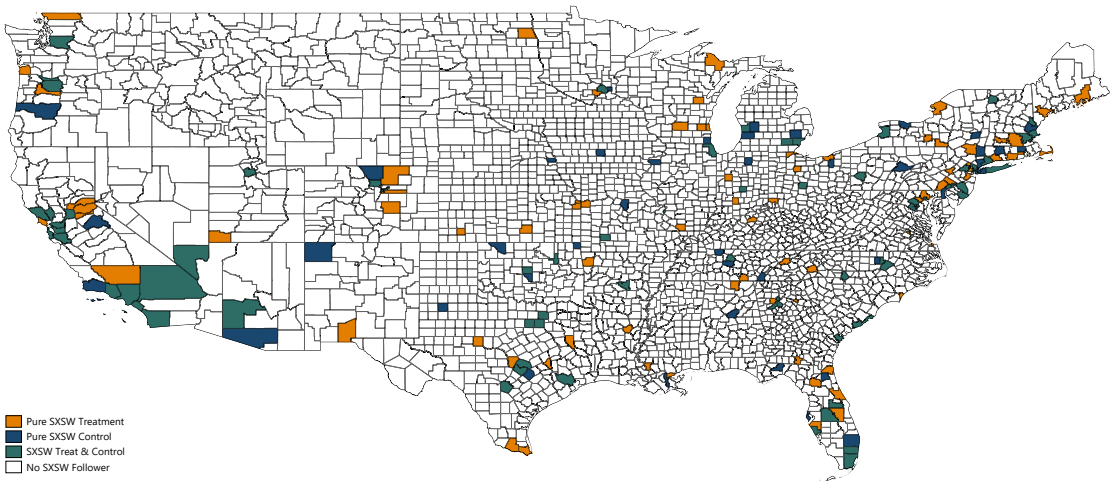
As a final step, we divide the rescaled values by their maximum and multiply them by 100. This is to re-normalize the Google trend values to take on values between 0 and 100.

Sources for Trump's golf activity

Table 2.17: Sources for Golf Data

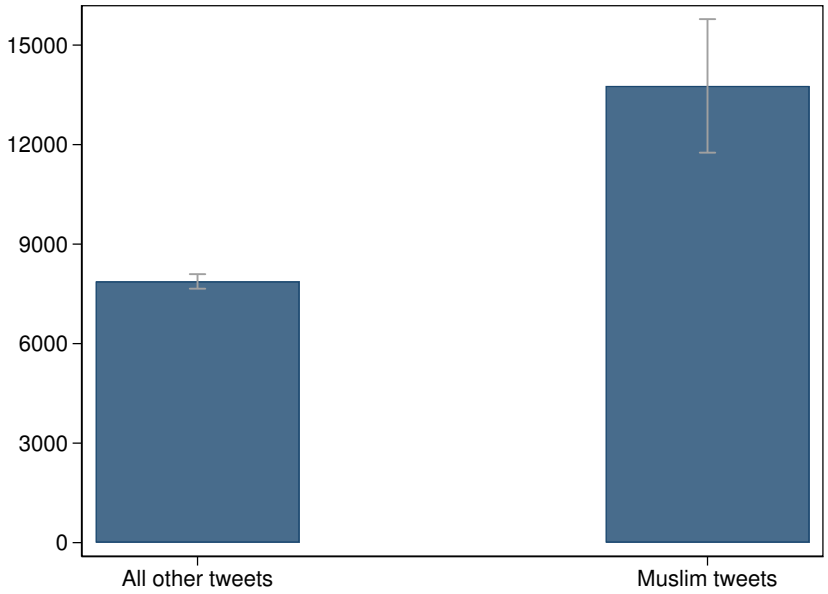
Source	Description
New York Times	The NYT tracks visits by Trump to his own properties. The data also track how often Trump visited a golf club.
trumpgolfcount.com	This website lists Trump's visits to golf clubs since his inauguration. It also provides additional analysis during which visits Trump likely played golf.
Presidential Schedule	The presidential schedule lists all past presidential journeys.

Figure 2.13: Identifying Variation



Notes: This map plots counties with SXSW followers who joined Twitter in March 2007 in orange; counties with SXSW followers who joined prior to the 2007 event in blue; and counties in both categories in green.

Figure 2.14: Average Retweets of Trump's Tweets, by Muslim Content

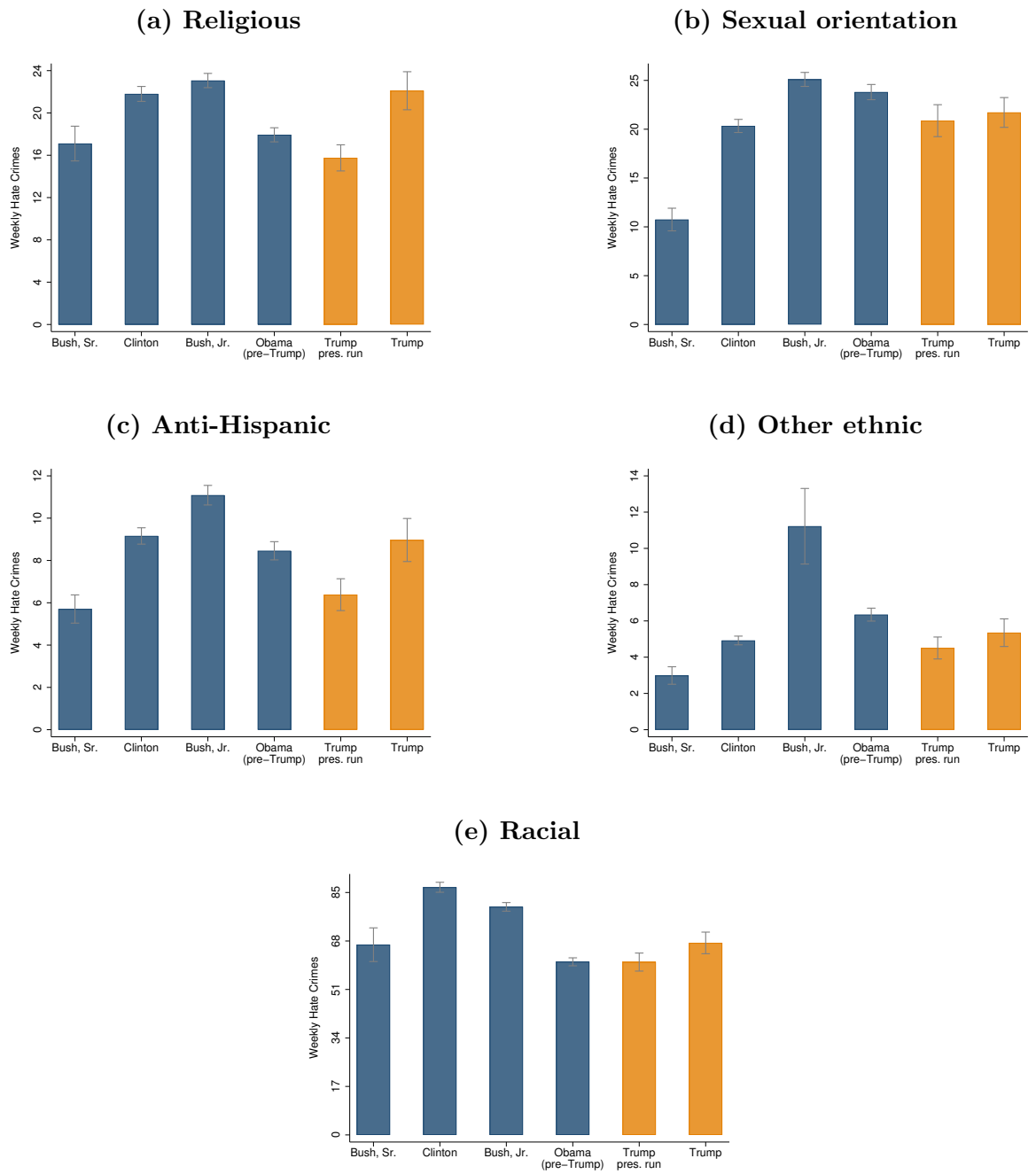


Notes: This figure plots the average number of retweets Donald Trump received on his tweets about Muslims compared to all other tweets. We also show 95% confidence intervals.

2.9.2 Appendix 2: Details on Trends in Hate Crimes by President

In this section, we provide some additional evidence on time series trends in hate crimes across US presidencies since 1990. A potential issue with the hate crime numbers we presented in Figure 2.1 might be that we consider all hate crimes jointly, which could hide underlying heterogeneous hate crime trends across groups. We thus reproduce the bar graphs using the other main categories of hate crimes in the FBI data (see Figure 2.15). Overall, the results yield a qualitatively similar conclusion. Trump does not appear to be an outlier for any of the main categories except Muslims.

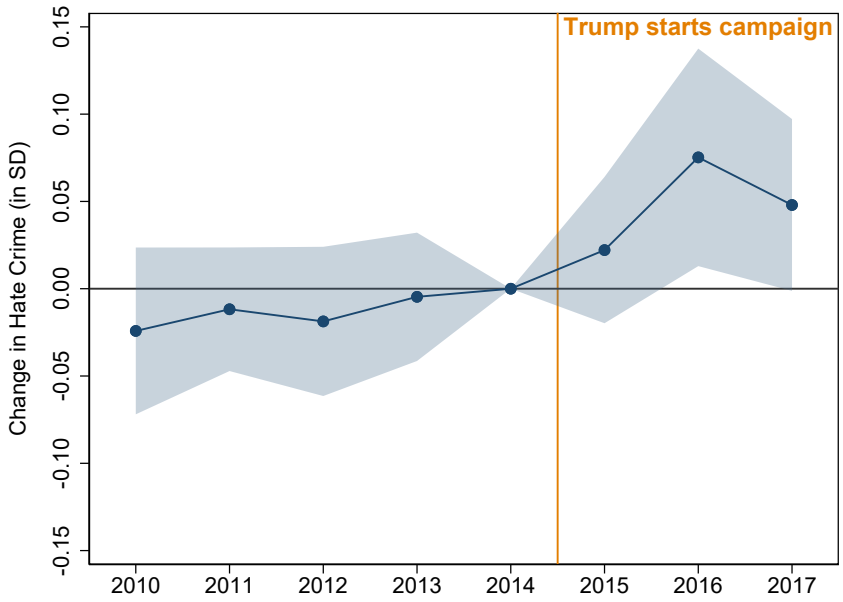
Figure 2.15: Average Weekly Hate Crimes since 1990, by President and Motivating Bias



Notes: This figure plots the average weekly number of hate crimes, by president and type of hate crime (as defined by the FBI). The headings indicate which type of hate crime is plotted. The whiskers indicate the 95% confidence intervals.

2.9.3 Appendix 3: Additional Cross-sectional Evidence

Figure 2.16: Change in Anti-Muslim Hate Crimes by Twitter Usage (Reduced Form)



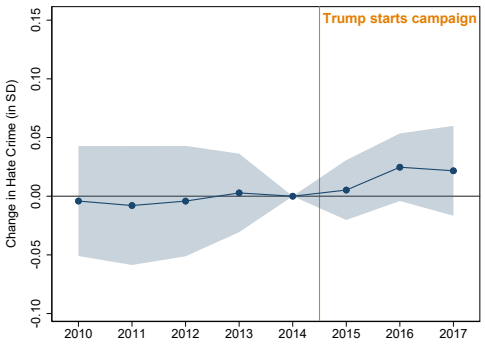
Notes: This figure plots the coefficients from running panel event study regressions as in Equation (2.3), where $\log(Twitter\ Usage)$ is replaced by $\log(SXSW\ followers, March\ 2007$. The dependent variable is the log number of hate crimes in a county. We standardized the variables to have a mean of zero and standard deviation of one. The vertical line indicates the start of Trump's presidential campaign start. The shaded areas are 95% confidence intervals.

Table 2.18: Descriptive Statistics (Main Variables)

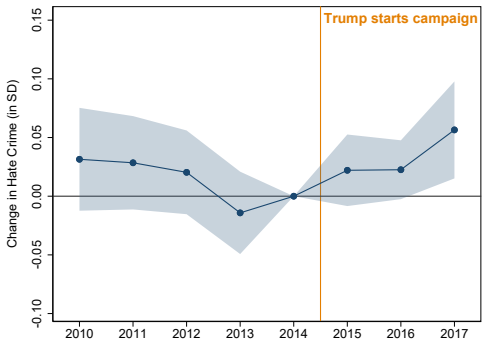
	Mean	Std. Dev.	Min.	Median	Max.	N
Hate crime and Twitter variables						
Δ Log(Hate crimes against Muslims)	0.02	0.13	-0.71	0.00	1.26	3108
Log(Twitter usage)	10.03	1.91	3.33	9.94	16.90	3108
Log(SXSW followers, March 2007)	0.06	0.32	0.00	0.00	4.98	3108
Log(SXSW followers, Pre)	0.02	0.18	0.00	0.00	3.61	3108
Demographic controls						
% aged 20-24	0.06	0.02	0.01	0.06	0.27	3108
% aged 25-29	0.06	0.01	0.03	0.06	0.15	3108
% aged 30-34	0.06	0.01	0.03	0.06	0.12	3108
% aged 35-39	0.06	0.01	0.03	0.06	0.11	3108
% aged 40-44	0.06	0.01	0.02	0.06	0.10	3108
% aged 45-49	0.06	0.01	0.02	0.06	0.09	3108
% aged 50+	0.39	0.07	0.11	0.39	0.75	3108
Population growth, 2000-2016	0.06	0.18	-0.43	0.03	1.32	3108
Geographical controls						
Population density	261.27	1733.47	0.10	45.60	69468.40	3108
Log(County area)	6.53	0.86	0.69	6.47	9.91	3108
Distance from Austin, TX (in miles)	1450.64	612.61	5.04	1464.66	3098.88	3108
Race and religion controls						
% white	0.77	0.20	0.03	0.84	0.98	3108
% black	0.09	0.14	0.00	0.02	0.85	3108
% native American	0.02	0.06	0.00	0.00	0.90	3108
% Asian	0.01	0.02	0.00	0.01	0.37	3108
% Hispanic	0.09	0.14	0.01	0.04	0.96	3108
% Muslim	0.23	1.08	0.00	0.00	30.35	3108
Socioeconomic controls						
% below poverty level	16.74	6.58	1.40	16.00	53.30	3108
% unemployed	5.50	1.94	1.80	5.30	24.10	3108
Gini index	0.44	0.03	0.33	0.44	0.65	3108
% uninsured	13.32	5.28	1.80	12.80	49.00	3108
Log(Median household income)	10.72	0.24	9.87	10.71	11.72	3107
% employed in agriculture	0.01	0.03	0.00	0.00	0.58	3108
% employed in IT	0.01	0.01	0.00	0.01	0.21	3108
% employed in manufacturing	0.16	0.13	0.00	0.13	0.72	3108
% employed in nontradable sector	0.29	0.11	0.00	0.28	1.00	3108
% employed in construction/real estate	0.07	0.05	0.00	0.06	1.00	3108
% employed in utilities	0.04	0.05	0.00	0.03	1.00	3108
% employed in business services	0.16	0.07	0.00	0.15	0.95	3108
% employed in other services	0.25	0.10	0.00	0.24	1.00	3108
% adults with high school degree	34.77	7.07	7.50	35.20	54.80	3108
% adults with graduate degree	7.05	4.12	0.00	5.80	44.40	3108

Figure 2.17: Change in Other Hate Crimes, by Twitter Usage (OLS)

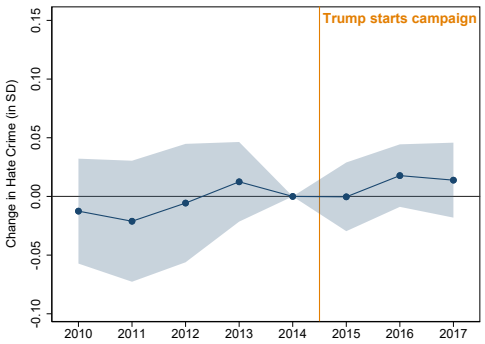
(a) All Hate Crimes



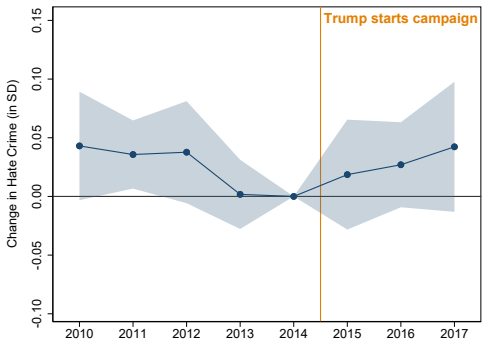
(b) Religious Hate Crimes



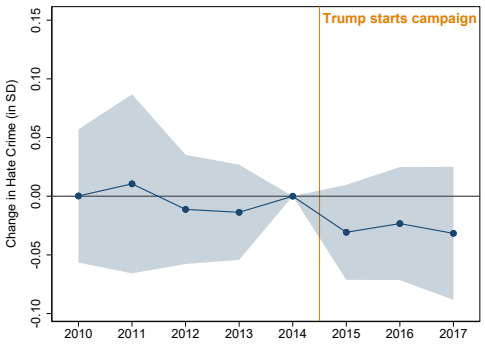
(c) Racial Hate Crimes



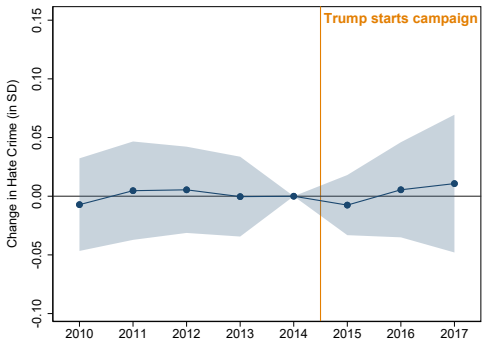
(d) Anti-Hispanic Hate Crimes



(e) Other Ethnic Hate Crimes



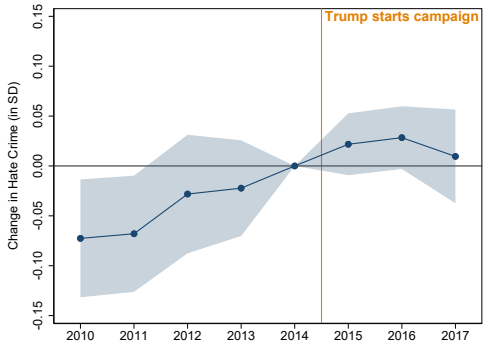
(f) Sexual Orientation Hate Crimes



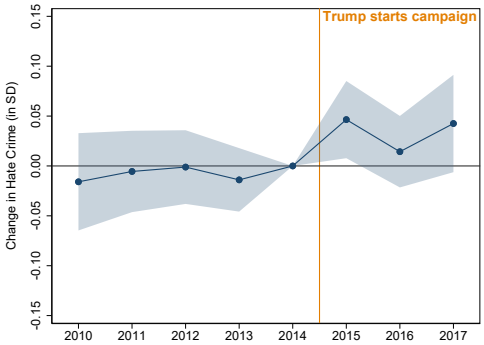
Notes: These figures plot the coefficients of running panel event study regressions as in Equation (2.3) for different types of hate crimes. We standardized the variables to have a mean of zero and standard deviation of one. The vertical line indicates the start of Trump's presidential campaign. The shaded areas are 95% confidence intervals. The excluded category is the year 2014.

Figure 2.18: Change in Other Hate Crimes, by Twitter Usage (Reduced Form)

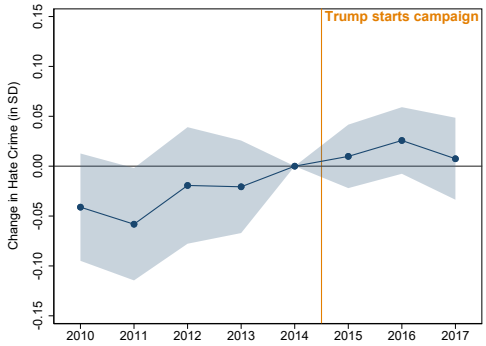
(a) All Hate Crimes



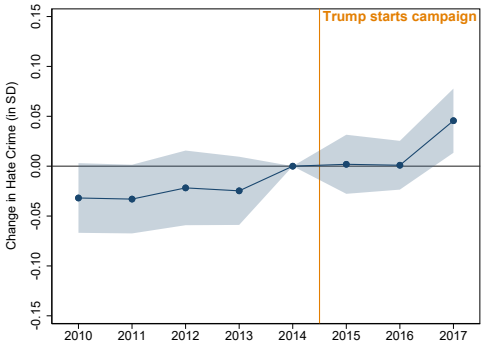
(b) Religious Hate Crimes



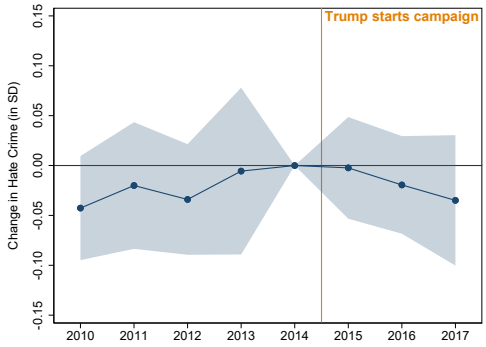
(c) Racial Hate Crimes



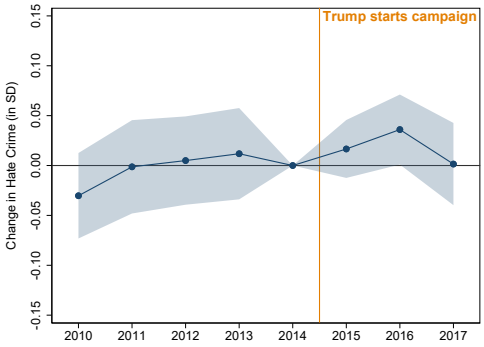
(d) Anti-Hispanic Hate Crimes



(e) Other Ethnic Hate Crimes



(f) Sexual Orientation Hate Crimes

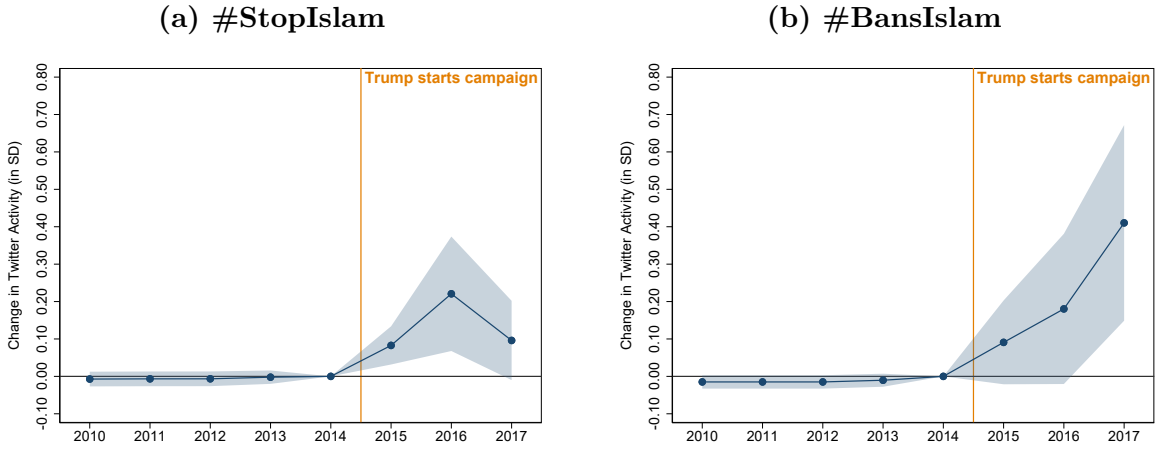


Notes: These figures plot the coefficients of running panel event study regressions as in Equation (2.3) for different types of hate crimes, where $\log(Twitter\ usage)$ is replaced with $\log(SXSW\ followers, March\ 2007$. We standardized the variables to have a mean of zero and standard deviation of one. The vertical line indicates the start of Trump's presidential campaign. The shaded areas are 95% confidence intervals. The excluded category is the year 2014.

Table 2.19: Descriptive Statistics (Main Variables, Continued)

	Mean	Std. Dev.	Min.	Median	Max.	N
Media controls						
% watching Fox News	0.26	0.01	0.23	0.26	0.30	3107
% watching prime time TV	0.43	0.01	0.40	0.43	0.47	3107
Election control						
Republican vote share, 2012	0.60	0.15	0.06	0.61	0.96	3108
Crime controls						
Violent crime rate	0.00	0.00	0.00	0.00	0.02	3108
Property crime rate	0.02	0.01	0.00	0.01	0.10	3108
Other hate crime variables						
Δ Log(Total hate crimes)	-0.01	0.36	-2.28	0.00	2.04	3108
Δ Log(Hate crimes against Hispanics)	-0.01	0.17	-1.65	0.00	1.21	3108
Δ Log(Other ethnicity-based hate crimes)	-0.02	0.16	-2.60	0.00	1.09	3108
Δ Log(Racially motivated hate crimes)	-0.01	0.31	-1.69	0.00	1.74	3108
Δ Log(Hate crimes based on sexual orientation)	-0.03	0.22	-1.46	0.00	1.20	3108
Δ Log(Hate crimes against other religions)	0.00	0.21	-1.58	0.00	1.59	3108
Log(Total hate crimes, ADL data)	0.23	0.63	0.00	0.00	5.38	3108

Figure 2.19: Change in Anti-Muslim Tweets (Reduced Form)



Notes: These figures plot the coefficients of running panel event study regressions as in Equation (2.3). The dependent variables are the log number of tweets containing the terms #BanIslam in panel (a) and #StopIslam in panel (b). We standardized the variables to have a mean of zero and standard deviation of one. The vertical line indicates the start of Trump's presidential campaign. The shaded areas are 95% confidence intervals. The excluded category is the year 2014.

Table 2.20: Comparing Counties with SXSW Followers, March 2007 vs. Pre

	March 2007 *and* Pre (1)	**March 2007 *only*** (2)	**Pre *only*** (3)	**Difference in means** (2) - (3)	t-stat
Demographic controls					
% aged 20-24	0.07	0.08	0.08	0.00	0.13
% aged 25-29	0.09	0.07	0.07	-0.00	-0.57
% aged 30-34	0.08	0.07	0.07	-0.00	-0.45
% aged 35-39	0.07	0.06	0.06	-0.00	-0.21
% aged 40-44	0.06	0.06	0.06	0.00	0.25
% aged 45-49	0.07	0.06	0.06	0.00	0.14
% aged 50+	0.32	0.35	0.35	-0.00	-0.03
Population growth, 2000-2016	0.18	0.18	0.15	0.03	0.67
Race and religion controls					
% white	0.50	0.65	0.67	-0.02	-0.53
% black	0.18	0.12	0.08	0.04	2.04**
% native American	0.01	0.01	0.02	-0.02	-1.03
% Asian	0.10	0.05	0.05	-0.01	-0.44
% Hispanic	0.20	0.16	0.15	0.01	0.32
% Muslim	1.31	0.81	0.75	0.05	0.20
Socioeconomic controls					
% below poverty level	15.71	15.82	13.69	2.14	1.94*
% unemployed	4.86	5.05	4.51	0.54	1.76*
Gini index	0.48	0.46	0.45	0.01	1.22
% uninsured	12.87	12.40	11.21	1.19	1.08
Log(Median household income)	11.00	10.91	10.99	-0.09	-1.57
% employed in agriculture	0.00	0.00	0.00	0.00	1.99∗
% employed in IT	0.04	0.02	0.02	-0.00	-0.02
% employed in manufacturing	0.07	0.09	0.09	0.01	0.55
% employed in nontradable sector	0.23	0.26	0.27	-0.01	-0.62
% employed in construction/real estate	0.06	0.07	0.07	0.01	1.02
% employed in utilities	0.04	0.04	0.03	0.00	0.53
% employed in business services	0.29	0.25	0.24	0.01	0.35
% employed in other services	0.27	0.26	0.28	-0.02	-0.94
% adults with high school degree	21.76	25.99	25.77	0.22	0.13
% adults with graduate degree	16.15	13.08	14.34	-1.26	-0.64
Media controls					
% watching Fox News	0.25	0.26	0.26	-0.00	-0.13
% watching prime time TV	0.42	0.43	0.43	0.00	0.11
Election control					
Republican vote share, 2012	0.33	0.46	0.47	-0.02	-0.43
Crime controls					
Violent crime rate	0.01	0.00	0.00	0.00	0.02
Property crime rate	0.03	0.02	0.02	0.00	1.09
Geographical controls					
Population density	5192.27	1021.39	1998.35	-976.96	-0.91
Log(County area)	6.30	6.63	6.54	0.09	0.31
Distance from Austin, TX (in miles)	1775.99	1749.38	1626.64	122.74	0.68

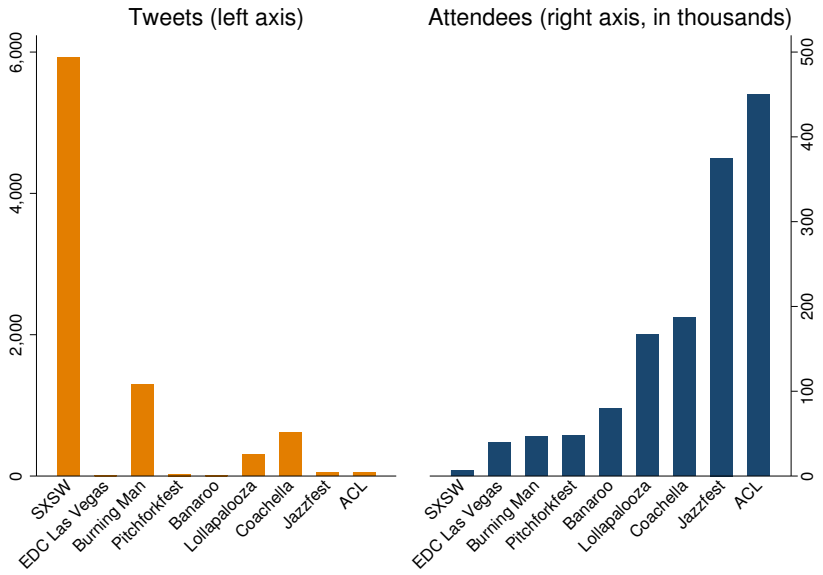
Notes: This table plots the mean values of the control variables for the three types of counties relevant for the cross-sectional results: (1) counties with new SXSW followers in March 2007 *and* the pre-period; (2) counties with new SXSW followers in March 2007 but no new followers in the pre-period; and (3) counties with new SXSW followers in the pre-period but no new followers in March 2007. $t-stat$ reports the result from a simple t-test for the equality of means between the counties with the key identifying variation. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.21: Balancedness SXSW Counties Individual Characteristics

First names (Corr. = 0.69)		**Terms used in bio (Corr. = 0.92)**	
Pre-Period	Treatment Period	Pre-Period	Treatment Period
michael	michael	http	http
mike	john	founder	com
paul	chris	com	digital
chris	jeff	co	founder
ryan	matt	tech	medium
eric	brian	design	director
david	david	director	tech
matthew	alex	product	music
john	jason	digital	social
jeff	kevin	designer	marketing
robert	paul	medium	design
mark	mike	music	co
andrew	dan	social	writer
daniel	andrew	love	love
james	peter	marketing	lover
kevin	jim	web	dad
jay	tom	geek	creative
jonathan	jennifer	writer	tweet
rob	steve	technology	author
rachel	todd	dad	designer

Notes: This table plots the ranking of the most common first names and terms used in a Twitter user's "bio" among users who follow "South by Southwest" on Twitter, depending on whether they signed up during the SXSW 2007 event or in the pre-period.

Figure 2.20: Number of Tweets and Attendees for Different Festivals (Full Year)



Notes: This figure plots the number of tweets mentioning major festivals in 2007.

Table 2.22: Correlation of Log(Twitter Users) across Events

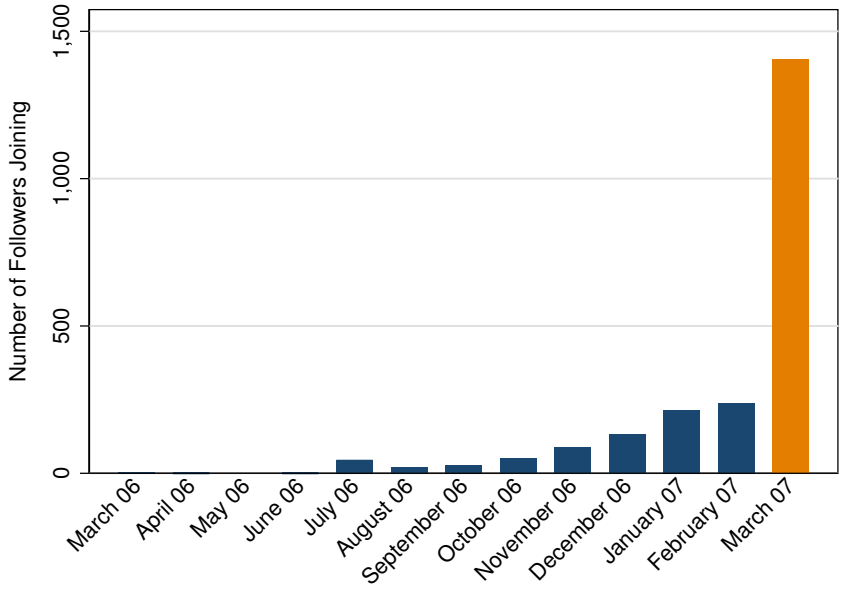
	SXSW March 2007	SXSW Pre	Coachella April 2007	Burning Man August 2007	Lollapalooza August 2007
SXSW followers, March 2007	1				
SXSW followers, Pre	0.77	1			
Coachella users, April 2007	0.44	0.48	1		
Burning Man users, August 2007	0.52	0.56	0.54	1	
Lollapalooza users, August 2007	0.03	0.06	0.00	0.00	1

Notes: This table reports the Pearson correlation coefficients between the main measure of interest (*SXSW followers, March 2007*) and different control variables. "Followers" are based on the locations of people who started following SXSW in a given month; "users" are based on people who tweeted at least once about a festival. We take the natural logarithm of these numbers with one added inside.

Table 2.23: Number of Counties With Any Twitter Users at SXSW or Other Festivals

	SXSW March 2007	SXSW Pre	Coachella April 2007	Burning Man August 2007	Lollapalooza August 2007
No followers	2953	2987	3091	3098	3105
At least 1 follower	155	121	17	10	3

Figure 2.21: Number of SXSW Followers Joining Each Month



Notes: This figure plots the number of SXSW followers who joined Twitter each month running up to the 2007 SXSW Festival. The orange bar marks the main instrument used in the paper.

Table 2.24: Robustness - Twitter Penetration Controls Based on Other Festivals in 2007

	ΔLog(Hate crimes against Muslims)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: OLS - Hate crimes against Muslims								
Log(Twitter usage)	0.021***	0.019***	0.019***	0.015***	0.015***	0.016***	0.015***	0.015***
	(0.006)	(0.006)	(0.007)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)
Panel B: Reduced form - Hate crimes against Muslims								
Log(SXSW followers, March 2007)	0.081***	0.081***	0.091***	0.080***	0.076***	0.076***	0.076***	0.076***
	(0.025)	(0.024)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
Panel C: 2SLS - Hate crimes against Muslims								
Log(Twitter usage)	0.153***	0.168***	0.198***	0.189***	0.187***	0.194***	0.205***	0.210***
	(0.045)	(0.046)	(0.049)	(0.055)	(0.055)	(0.057)	(0.063)	(0.064)
Weak IV 95% AR confidence set	[0.06; 0.23]	[0.08; 0.25]	[0 .1; 0.28]	[0.08; 0.29]	[0.08; 0.28]	[0.08; 0.29]	[0.08; 0.33]	[0.09; 0.34]
Log(Burning Man users, August 2007)	-0.003	-0.003	0.031	-0.016	-0.021	-0.020	-0.009	0.008
	(0.083)	(0.084)	(0.084)	(0.078)	(0.077)	(0.077)	(0.075)	(0.072)
Log(Coachella users, April 2007)	0.007	-0.005	0.018	0.002	-0.005	-0.011	-0.014	-0.020
	(0.100)	(0.106)	(0.117)	(0.108)	(0.111)	(0.112)	(0.112)	(0.112)
Log(Lollapalooza users, August 2007)	0.263	0.261	0.251	0.243	0.240	0.241	0.240	0.234
	(0.187)	(0.198)	(0.194)	(0.197)	(0.191)	(0.196)	(0.194)	(0.194)
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Population controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Race and religion controls				Yes	Yes	Yes	Yes	Yes
Socioeconomic controls					Yes	Yes	Yes	Yes
Media controls						Yes	Yes	Yes
Election control							Yes	Yes
Crime controls								Yes
Geographical controls			Yes	Yes	Yes	Yes	Yes	Yes
Observations	3107	3107	3107	3107	3106	3105	3105	3105
Mean of DV	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019
Robust F-stat.	158.32	117.39	101.21	112.09	87.87	80.59	67.50	66.25

Notes: This table presents county-level OLS, first stage, and IV regressions where the dependent variable is the log change in hate crimes against Muslims between 2010 and 2017. *Log(Twitter usage)* is instrumented using the number of users who started following SXSW in March 2007. The other variables count the number of users tweeting about any of the three largest US music festivals in 2007: Cochella, Burning Man, and Lollapalooza. All regressions control for population deciles, state fixed effects, and the full set of controls as in column 8 of Table 2.3 (not shown). Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) using the Stata package from Sun (2018). For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.25: Robustness - Alternative Measures of Twitter Usage

	Survey # households using Twitter (1)	Survey % households using Twitter (2)	GESIS Tweets (Pre-Trump) (3)	GESIS Twitter users (4)
Panel A: OLS - Hate crimes against Muslims				
Twitter usage measure	0.059***	0.024**	0.017***	0.003**
	(0.020)	(0.010)	(0.006)	(0.001)
Panel B: First stage - Twitter usage				
Log(SXSW followers, March 2007)	0.440***	0.080***	0.443***	0.634***
	(0.041)	(0.018)	(0.061)	(0.157)
Panel C: 2SLS - Hate crimes against Muslims				
Twitter usage measure	0.169**	0.926**	0.167**	0.117**
	(0.067)	(0.387)	(0.072)	(0.057)
Weak IV 95% AR confidence set	[0.04; 0.29]	[0.28; 1.87]	[0.04; 0.31]	[0.03; 0.27]
Log(SXSW followers, Pre)	0.014	-0.021	0.008	-0.014
	(0.062)	(0.090)	(0.070)	(0.077)
Observations	3106	3106	3107	3107
Mean of DV	0.019	0.019	0.019	0.019
Robust F-stat.	114.10	20.59	53.15	16.35

Notes: This table presents county-level OLS, reduced form, and IV regressions where the dependent variable is the log change in hate crimes against Muslims between 2010 and 2017. *Twitter usage measure* is the measure listed in the top row, instrumented using the number of users who started following SXSW in March 2007 (in log with 1 added inside). *SXSW followers, Pre* is the number of SXSW followers who registered at some point in 2006 (in log with 1 added inside). All regressions control for population deciles and state fixed effects, as well as demographic controls including population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) using the Stata package from Sun (2018). For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.26: 2SLS - Alternative SXSW Controls

SXSW measure	Followers	**Followers**	Followers	Followers	Followers	Followers	Tweets	Tweets	Tweets	Tweets	Tweets
Transformation	Log	**Log**	Log	Log	Log	Log	Dummies	Dummies	Dummies	Dummies	Dummies
Control variable(s)	None	**Pooled**	Pooled	Individual	Individual	Individual	Pooled	Pooled	Individual	Individual	Individual
Control period	–	**2006**	2006-Feb. 2007	Feb. 2007	2006	2006-Feb. 2007	2006	2006-Feb. 2007	Feb. 2007	2006	2006-Feb. 2007
Control counties	–	**67**	121	59	67	121	109	154	55	109	154
Corr(March 2007, Control), average	–	**0.77**	0.83	0.72	0.49	0.54	0.45	0.44	0.42	0.28	0.30
	(1)	**(2)**	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Panel A: Reduced form - Hate Crimes against Muslims											
SXSW measure, March 2007	0.089***	**0.074****	0.077**	0.088**	0.071**	0.070**	0.092***	0.103***	0.090***	0.064**	0.066**
	(0.021)	**(0.030)**	(0.037)	(0.033)	(0.028)	(0.035)	(0.026)	(0.026)	(0.028)	(0.027)	(0.027)
SXSW measure, control (linear combination)	–	**0.034**	0.019	0.002	-0.241	-0.237	0.032	-0.008	0.058	-0.106	-0.117
		(0.054)	(0.041)	(0.053)	(0.246)	(0.247)	(0.031)	(0.025)	(0.050)	(0.095)	0.107
Panel B: 2SLS - Hate Crimes against Muslims											
Log(Twitter usage)	0.167***	**0.161****	0.272**	0.189***	0.155**	0.204**	0.319***	0.344***	0.344***	0.297**	0.362**
	(0.036)	**(0.069)**	(0.131)	(0.061)	(0.062)	(0.099)	(0.099)	(0.102)	(0.102)	(0.131)	(0.173)
Weak IV 95% AR confidence set	[0.10; 0.23]	**[0.05; 0.30]**	[0.03; 0.59]	[0.06; 0.30]	[0.04; 0.28]	[0.02; 0.41]	[0.16; 0.56]	[0.18; 0.59]	[0.18; 0.59]	[0.08; 0.72]	[0.11; 1.15]
Observations	3,107	**3,107**	3,107	3,107	3,107	3,107	3,105	3,105	3,105	3,105	3,105
Robust F-stat.	165.7	**58.04**	16.67	48.02	76.74	34.02	24.34	26.59	26.59	10.63	7.257

Notes: This table presents county-level OLS and IV regressions where the dependent variable is the log change in hate crimes against Muslims between 2010 and 2017. *Log(Twitter usage)* is instrumented using the measure described in the top rows; column 2 plots the baseline specification. *SXSW measure, control (linear combination)* is the estimate for the SXSW control variable. "Pooled" controls refer to one variable for the entire control period; "individual" to a vector of individual variables for each control period (e.g. one variable for March 2006, one variable for April 2006, etc.). For the case of individual controls, we plot the linear combinations of the coefficients and associated standard errors. In those cases, we also plot the *average* of the correlation of the individual controls with the March 2007 measure. "Followers" are based on Twitter users in a county that started following SXSW in a given month. "Tweets" are based on whether we can identify any user that tweeted about SXSW in a given month. All regressions control for population deciles, state fixed effects and demographic controls that include population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. The specifications in columns 7 through 11 include the full vector of control variables. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) using the Stata package from Sun (2018). For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.27: Social Media and Types of Hate Crimes

	Any (1)	Vandalism (2)	Theft (3)	Burglary (4)	Robbery (5)	Assault (6)
Panel A: OLS - Hate crimes against Muslims						
Log(Twitter usage)	0.019***	0.008	0.001*	0.001	0.001	0.018***
	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)	(0.006)
Panel B: Reduced form - Hate crimes against Muslims						
Log(SXSW followers, March 2007)	0.074**	0.031	0.003	0.007	0.000	0.067**
	(0.030)	(0.022)	(0.005)	(0.010)	(0.004)	(0.029)
Panel C: 2SLS - Hate crimes against Muslims						
Log(Twitter usage)	0.161**	0.068	0.007	0.014	0.001	0.146**
	(0.069)	(0.047)	(0.011)	(0.021)	(0.008)	(0.066)
Weak IV 95% AR confidence set	[0.04; 0.30]	[0.01; 0.15]	[0.01; 0.03]	[0.02; 0.05]	[0.01; 0.01]	[0.03; 0.28]
Log(SXSW followers, Pre)	0.008	0.036	-0.004	-0.016	0.017	0.016
	(0.069)	(0.051)	(0.008)	(0.017)	(0.021)	(0.060)
Observations	3107	3107	3107	3107	3107	3107
Mean of DV	0.019	0.008	0.000	0.000	0.001	0.014
Robust F-stat.	58.04	58.04	58.04	58.04	58.04	58.04

Notes: This table presents county-level OLS and IV regressions where the dependent variable is the log change in hate crimes against Muslims of the type in the top row between 2010 and 2017. *Log(Twitter usage)* is instrumented using the number of users who started following SXSW in March 2007. *SXSW followers, Pre* is the number of SXSW followers who registered at some point in 2006. All regressions control for population deciles and state fixed effects (not shown). Demographic controls include population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. Race and religion controls contains the share of people identifying as white, African American, Native American or Pacific Islander, Asian, Hispanic, or Muslim. Socioeconomic controls include the poverty rate, unemployment rate, local GINI index, the share of uninsured individuals, log median household income, the share of highschool graduates, the share of people with a graduate degree, as well as the employment shares in agriculture, information technology, manufacturing, nontradables, construction and real estate, utilities, business services, or other sectors. Media controls include the viewership share of Fox News, the cable TV spending to population ratio, and the prime time TV viewership to population ratio. Election control is the county-level vote share of the Republican party in 2012. Crime controls are the rates of violent or property crime from the FBI. Geographical controls include the linear distance from the SXSW festival location (Austin, Texas), population density, and the natural logarithm of county size. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) using the Stata package from Sun (2018). For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.28: Social Media and Hate Crimes – Alternative Standard Errors

	Robust SE (1)	Bootstrap robust SE (2)	Bootstrap state cluster SE (3)	Spatial SE (4)
Panel A: OLS - Hate crimes against Muslims				
Log(Twitter usage)	0.019***	0.019***	0.019***	0.019***
	(0.005)	(0.005)	(0.006)	(0.005)
Panel B: Reduced form - Hate crimes against Muslims				
Log(SXSW followers, March 2007)	0.074***	0.074**	0.074***	0.074***
	(0.029)	(0.031)	(0.027)	(0.028)
Panel C: 2SLS - Hate crimes against Muslims				
Log(Twitter usage)	0.161**	0.161**	0.161**	0.161**
	(0.066)	(0.069)	(0.071)	(0.067)
Weak IV 95% AR confidence set	[0.05; 0.30]			
Log(SXSW followers, Pre)	0.008	0.008	0.008	0.008
	(0.057)	(0.057)	(0.077)	(0.064)
Observations	3107	3107	3107	3107
Mean of DV	0.019	0.019	0.019	0.019
Robust F-stat.	39.37	39.37	57.15	52.14

Notes: This table presents county-level OLS and IV regressions where the dependent variable is the log change in hate crimes against Muslims between 2010 and 2017. *Log(Twitter usage)* is instrumented using the number of users who started following SXSW in March 2007. *SXSW followers, Pre* is the number of SXSW followers who registered at some point in 2006. All regressions control for population deciles and state fixed effects (not shown). Demographic controls include population growth between 2000 and 2016 as well as age cohort controls for the share of people aged 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and those over 50. Spatial standard errors are based on the method proposed in Colella et al. (2019), implemented in Stata as *acreg*, using a 200 miles cutoff. For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Standard errors are computed as indicated in the top row. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.29: Heterogeneous Effects – Hate Groups and Hate Crimes

Dependent variable: Log(Anti-Muslim hate crimes)	(1) No hate groups	(2) Any hate group	(3) Few hate crimes	(4) Many hate crimes
Panel A: OLS				
Log(Twitter Usage) x Year=2010	-0.01	-0.01	-0.00	-0.07
	(0.01)	(0.09)	(0.00)	(0.11)
Log(Twitter Usage) x Year=2011	-0.00	0.01	0.00	0.01
	(0.01)	(0.11)	(0.00)	(0.13)
Log(Twitter Usage) x Year=2012	-0.00	-0.02	0.00	-0.02
	(0.01)	(0.14)	(0.00)	(0.15)
Log(Twitter Usage) x Year=2013	-0.01	-0.00	0.00	-0.04
	(0.01)	(0.11)	(0.00)	(0.13)
Log(Twitter Usage) x Year=2015	0.01	0.45***	0.00	0.52***
	(0.01)	(0.14)	(0.00)	(0.15)
Log(Twitter Usage) x Year=2016	0.01	0.58***	0.01**	0.63***
	(0.01)	(0.17)	(0.00)	(0.18)
Log(Twitter Usage) x Year=2017	-0.01	0.38	0.00	0.34
	(0.01)	(0.23)	(0.00)	(0.25)
Panel B: Reduced form				
Log(SXSW followers) x Year=2010	-0.07**	-0.01	-0.00	-0.03
	(0.03)	(0.04)	(0.00)	(0.03)
Log(SXSW followers) x Year=2011	-0.04*	0.00	-0.00	0.00
	(0.02)	(0.03)	(0.00)	(0.03)
Log(SXSW followers) x Year=2012	-0.03	-0.02	0.00	-0.02
	(0.02)	(0.03)	(0.01)	(0.03)
Log(SXSW followers) x Year=2013	-0.05*	0.02	-0.00	0.01
	(0.03)	(0.03)	(0.00)	(0.03)
Log(SXSW followers) x Year=2015	-0.01	0.03	-0.00	0.10***
	(0.03)	(0.03)	(0.00)	(0.03)
Log(SXSW followers) x Year=2016	0.02	0.09*	-0.01	0.14***
	(0.03)	(0.05)	(0.01)	(0.04)
Log(SXSW followers) x Year=2017	-0.01	0.06*	-0.00	0.13***
	(0.03)	(0.03)	(0.01)	(0.05)
County FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Pop. deciles x Year FE	Yes	Yes	Yes	Yes
Observations	1145248	147680	1156896	136032

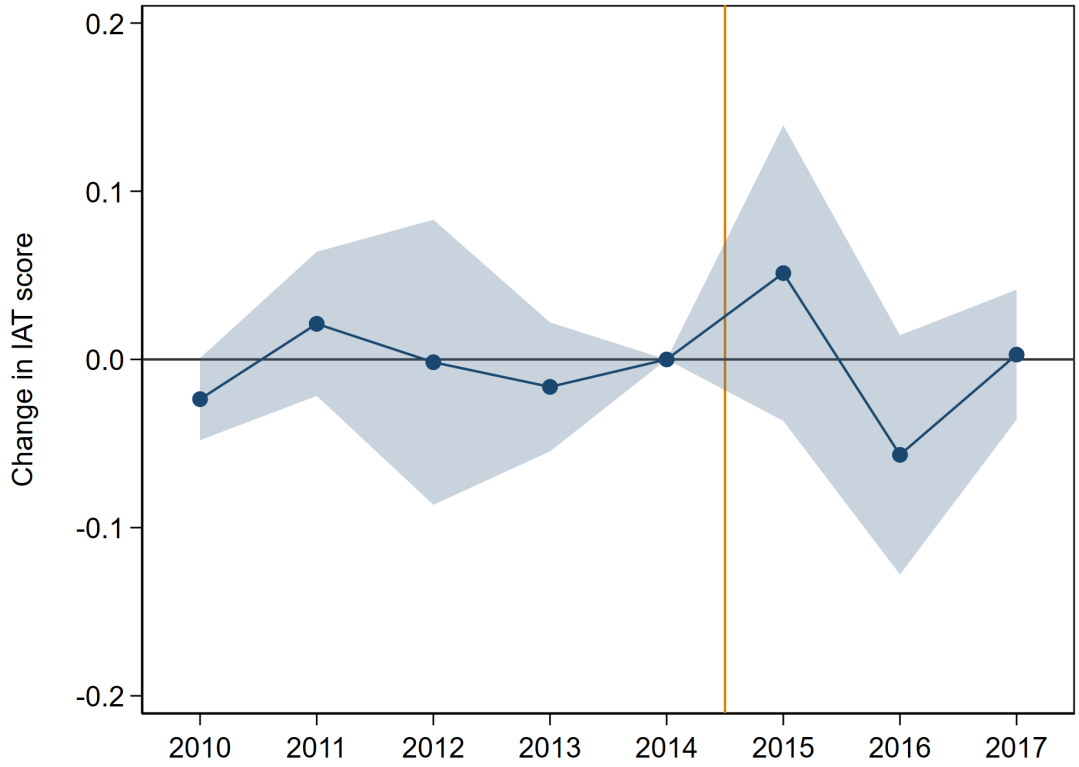
Notes: This table presents panel event study regressions where the dependent variable is the log number of hate crimes against Muslims (with one added inside). We standardized the variables to have a mean of zero and standard deviation of one. The sample period is 2010 to 2017. 2014 is the excluded period. $Log(SXSWfollowers)$ is the number of local SXSW followers that joined Twitter in March 2007. The existence of hate groups is based on data from the Southern Poverty Law Center (SPLC). The number of hate crimes in the pre-period is based on the total number of hate crimes per capita the FBI registered in a county from 2010 until 2015, split at the 90th percentile. All regressions control for the interaction of population deciles with year dummies. Standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.30: Social Media and Changes in Implicit Bias against Muslims

	Raw IAT scores (1)	Residual IAT scores (2)	Only conservatives (3)	Only whites (4)	Only Christians (5)	Only non-Muslims (6)	Only obligatory tests (7)	At least 10 tests (8)
Panel A: OLS - Change in implicit bias against Muslims								
Log(Twitter usage)	0.026*	0.023	-0.023	-0.012	0.002	0.021	0.017	0.004
	(0.014)	(0.014)	(0.022)	(0.014)	(0.012)	(0.014)	(0.021)	(0.006)
Panel B: Reduced form - Change in implicit bias against Muslims								
Log(SXSW followers, March 2007)	-0.016	-0.014	-0.023	-0.012	-0.027	-0.017	-0.003	0.006
	(0.016)	(0.014)	(0.035)	(0.022)	(0.020)	(0.016)	(0.022)	(0.008)
Panel C: 2SLS - Change in implicit bias against Muslims								
Log(Twitter usage)	-0.043	-0.039	-0.061	-0.035	-0.077	-0.048	-0.007	0.017
	(0.046)	(0.039)	(0.096)	(0.066)	(0.064)	(0.048)	(0.058)	(0.024)
Weak IV 95% AR confidence set	[-0.14; 0.03]	[-0.11; 0.02]	[-0.27; 0.11]	[-0.18; 0.07]	[-0.21; 0.02]	[-0.15; 0.03]	[-0.12; 0.09]	[-0.02; 0.06]
Log(SXSW followers, Pre)	0.024	0.011	-0.036	-0.027	0.040	0.020	-0.015	-0.001
	(0.019)	(0.017)	(0.051)	(0.029)	(0.028)	(0.019)	(0.025)	(0.017)
Observations	2251	2222	1303	1945	1987	2230	1759	571
Mean of DV	-0.038	-0.013	-0.007	-0.053	-0.032	-0.039	-0.039	-0.044
Robust F-stat.	49.42	51.78	34.80	36.57	36.79	38.15	64.48	28.31

Notes: This table presents county-level OLS and IV regressions where the dependent variable is the change in average Implicit Association Test (IAT) scores that measures implicit bias against Muslims between 2010 and 2017. Higher scores reflect more bias. *Log(Twitter usage)* is instrumented using the number of users who started following SXSW in March 2007. *SXSW followers, Pre* is the number of SXSW followers who registered at some point in 2006. All regressions control for population deciles and state fixed effects (not shown). In column 2, IAT scores are residualized with respect to age and its squared term, as well as a full set of fixed effects for educational attainment, race, sex, and ethnicity. In columns 3 through 6, the sample is restricted to respondents as indicated in the top row. Column 7 only includes tests that are obligatory, e.g. as part of a work program. Column 8 restricts the sample to counties with at least 10 IAT tests before and after Trump's presidential run. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) using the Stata package from Sun (2018). For the just-identified case we study here, the "robust" F-stat. is equivalent to the "Kleibergen-Paap" or the "effective" F-statistic of Olea and Pflueger (2013). Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

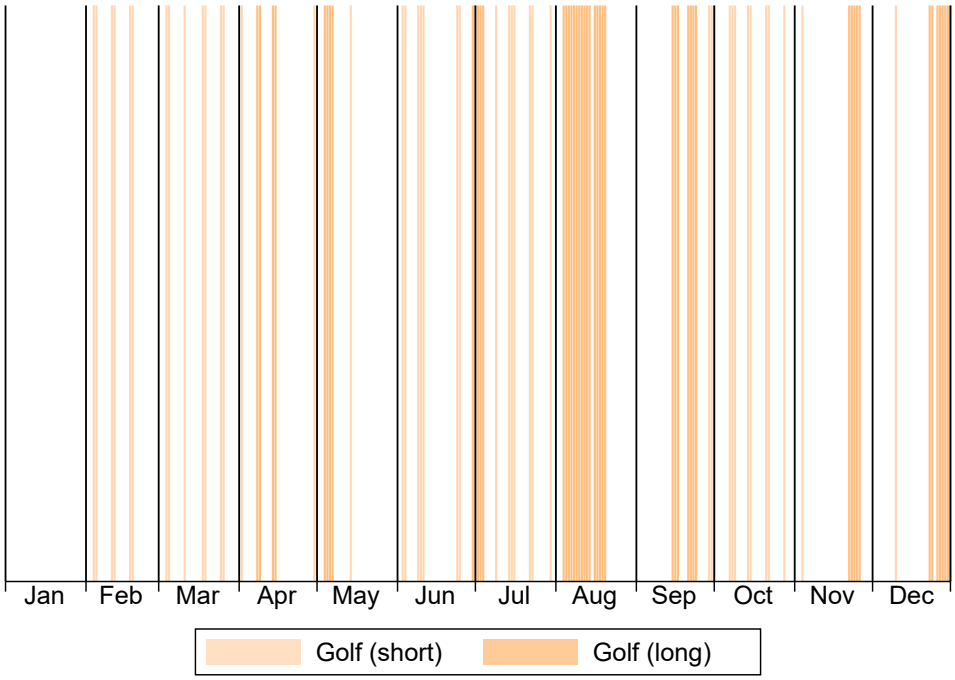
Figure 2.22: Change in Implicit Bias (Reduced Form)



Notes: These figures plot the coefficients of running a panel event study regression as in Equation (2.3). The dependent variable is the mean county-level IAT score that measures implicit bias against Muslims. We standardize the variables to have a mean of zero and standard deviation of one. The vertical line indicates the start of Trump's presidential campaign. The shaded areas are 95% confidence intervals. The excluded category is the year 2014.

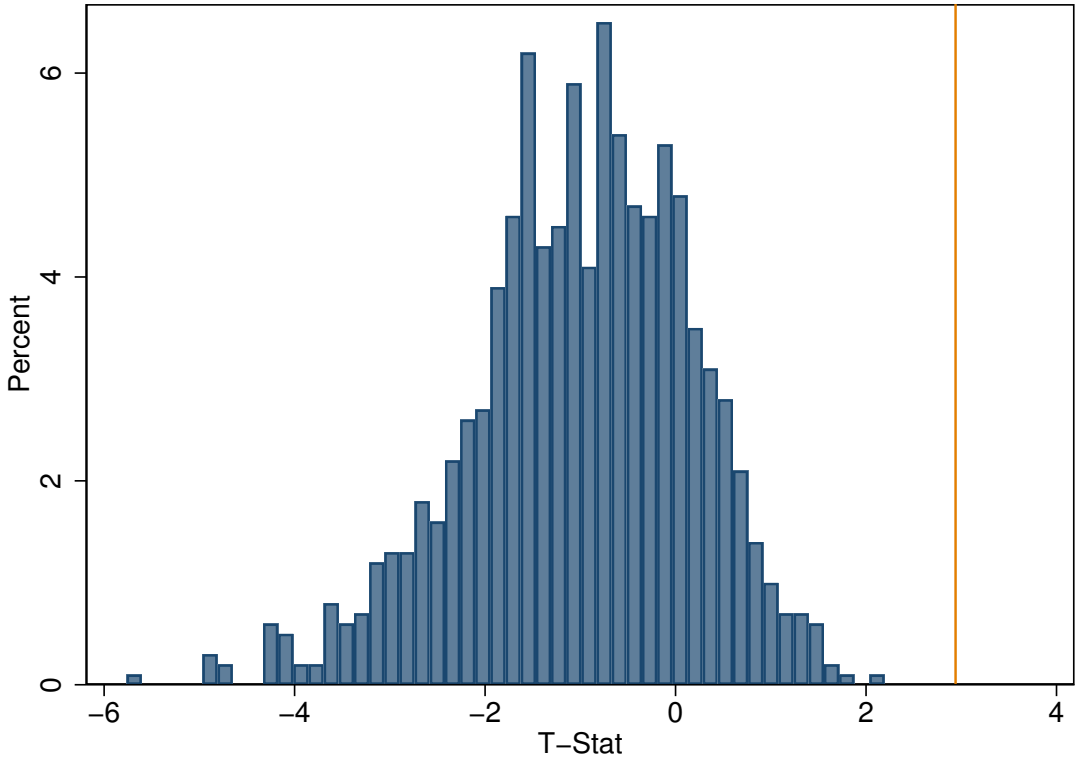
2.9.4 Appendix 4: Additional Time Series Evidence

Figure 2.23: Trump's Golf Days in 2017



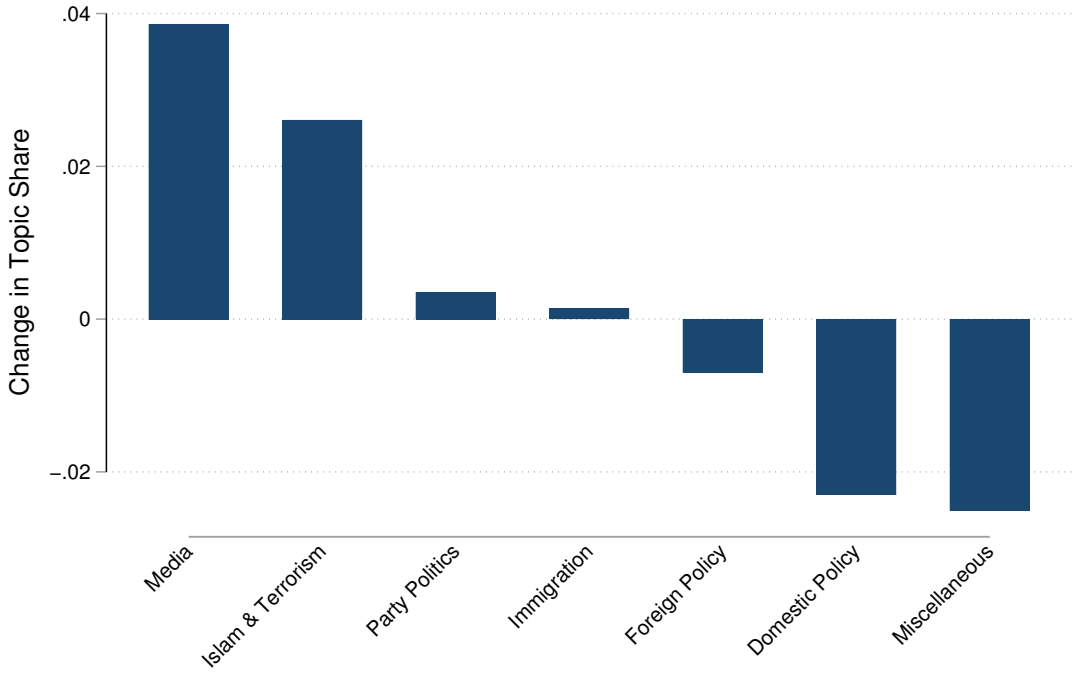
Notes: This figure plot the days in 2017 when Donald Trump played golf. Golf (long) indicates three or more consecutive days of golf.

Figure 2.24: Randomization Test for Golf Days



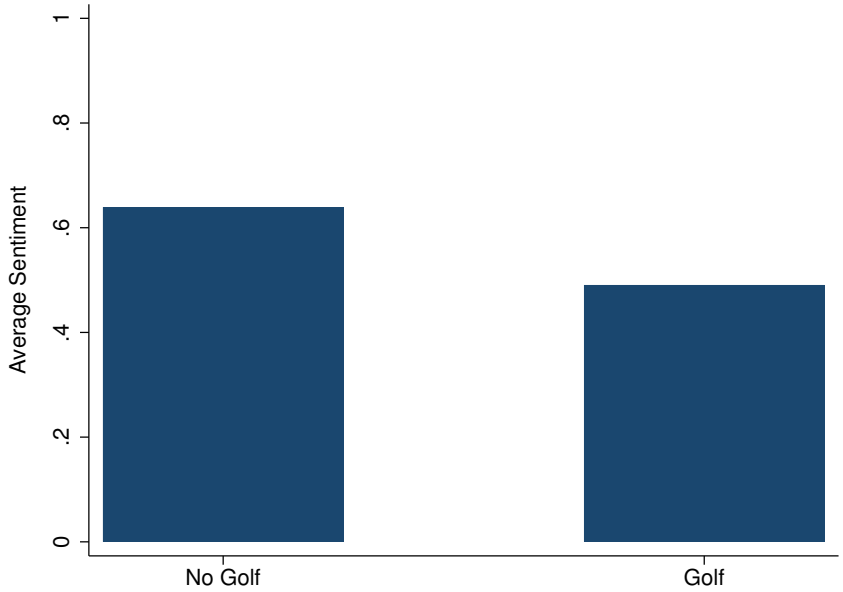
Notes: This figure visualizes the distribution of t-statistics from a randomization test of the first stage regression of Trump's tweets about Muslims on placebo golf days. In particular, we create 1,000 placebo sets of 92 golf days, which is the number of times Trump golfed in 2017. We then regress the log number of Trump's tweets about Muslims on these dummies using the baseline specification in Equation (2.6) and report the distribution of the resulting t-statistics. The orange line marks our baseline point estimate.

Figure 2.25: Shift in Topics of Trump's Tweets on Golf Days



Notes: This figure shows how the content of Trump's tweets changes on days when he plays golfs. These topics were hand-coded using Amazon Mechanical Turk.

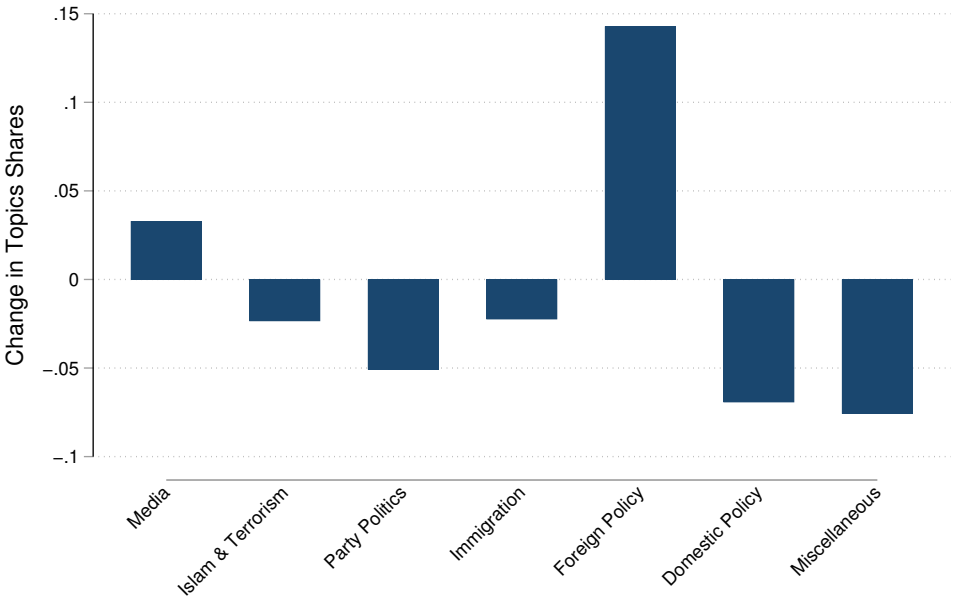
Figure 2.26: Trump's Tweets Are More Negative on Golf Days



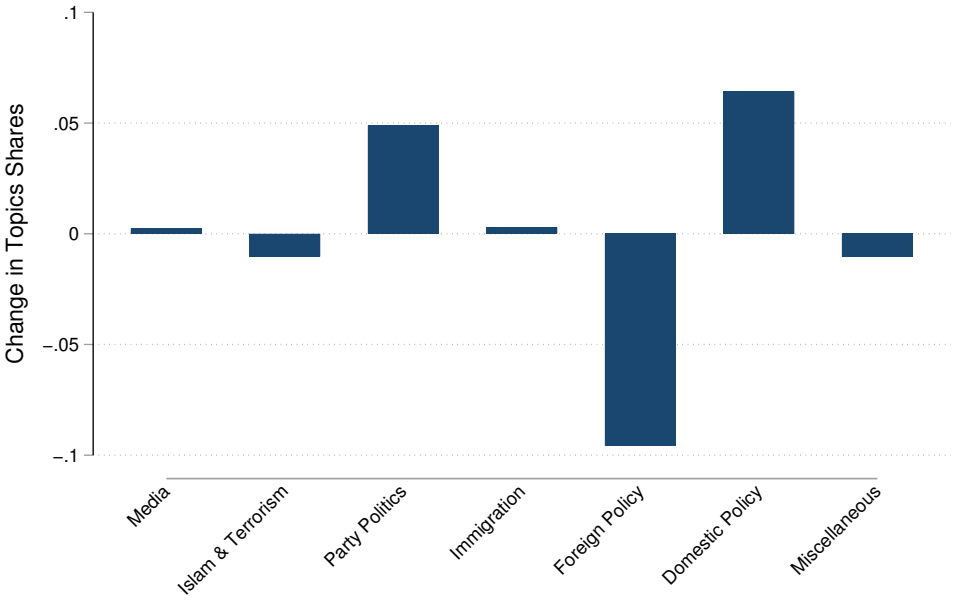
Notes: This figure plots the average sentiment of Trump's tweets on golf and non-golf days. Lower values mean more negative sentiment. The sentiment was hand-coded using Amazon Mechanical Turk on a scale from -2 to 2.

Figure 2.27: Shift in Topics of Trump's Tweets During Other Events

(a) Travel Abroad



(b) Policy Briefing



Notes: This figure shows how the content of Trump's tweets changes on days when he is traveling abroad (panel a) or receives a policy briefing (panel b). These topics were hand-coded using Amazon Mechanical Turk.

Table 2.31: Summary Statistics for Time Series

Variable	Mean	SD	p50	Min	Max	N
Trump tweets						
Muslim Trump tweets (1+log)	0.08	0.25	0.00	0.00	1.79	365
Total Trump tweets (1+log)	1.95	0.58	1.95	0.00	3.30	365
Muslim Trump tweets (dummy)	0.09	0.29	0.00	0.00	1.00	365
Hate crimes against Muslims (1 + natural logarithm)						
All types	0.45	0.47	0.69	0.00	1.79	365
Assault	0.31	0.42	0.00	0.00	1.61	365
Vandalism	0.15	0.30	0.00	0.00	1.39	365
Theft	0.01	0.09	0.00	0.00	1.10	365
Burglary	0.01	0.07	0.00	0.00	0.69	365
Robbery	0.01	0.09	0.00	0.00	0.69	365
Other hate crimes (1 + natural logarithm)						
All hate crimes	2.99	0.27	3.00	2.08	3.74	365
Ethnicity (incl. Hispanic)	0.44	0.47	0.69	0.00	1.79	365
Race	2.27	0.37	2.30	0.69	3.00	365
Sexual orientation	1.32	0.46	1.39	0.00	2.40	365
Religion (excl. Muslims)	1.40	0.50	1.39	0.00	2.89	365
TV news coverage (1 + natural logarithm)						
Muslim mentions (total)	3.71	0.64	3.69	0.69	5.26	365
Muslim mentions (Fox News)	2.75	0.66	2.77	0.00	4.29	365
Muslim mentions (CNN)	2.24	0.94	2.30	0.00	4.29	365
Muslim mentions (MSNBC)	2.75	0.66	2.77	0.00	4.26	365
Trump's golfing						
Trump golfs	0.25	0.43	0.00	0.00	1.00	365
Trump golfs (NYT only)	0.24	0.43	0.00	0.00	1.00	365
Trump golf (alternative)	0.25	0.44	0.00	0.00	1.00	365
Golf holiday	0.16	0.37	0.00	0.00	1.00	365
Golf at any point in previous week	0.71	0.45	1.00	0.00	1.00	365
Other control variables						
Google searches (PC)	-0.19	1.59	-0.48	-1.47	11.94	365
Terror attack in the US	0.00	0.05	0.00	0.00	1.00	365
Terror attack in Europe	0.03	0.17	0.00	0.00	1.00	365
Terror attack elsewhere	0.08	0.28	0.00	0.00	2.00	365

Notes: This table presents descriptive statistics for the IV sample. The sample year is 2017, for which we have information on Trump's golfing. *1+log* or *1+natural logarithm* means that the logarithm of any variable is calculated with 1 added inside. The data on hate crimes come from the FBI hate crime statistics. Data on Trump's golfing come from the New York Times, the official White House presidential schedule, and trumpgolfcount.com. *Google searches (PC)* is the first principal component of Google trends for the key words "islam", "mosque", "muslim", "refugee", "sharia", and "terror". We use these same keywords as measures of TV news attention based on data from the internet archive. The sources for the number of terror attacks is the Global Terrorism Database. See the online appendix for more details on data and variable construction.

Table 2.32: Summary Statistics by Day of Week (2017 only)

Day of week		Hate crimes against Muslims	Tweets about Muslims	Trump golfs
Monday	Sum	43	3	4
	Mean	0.83	0.06	0.08
Tuesday	Sum	33	6	3
	Mean	0.63	0.12	0.06
Wednesday	Sum	43	10	4
	Mean	0.83	0.19	0.08
Thursday	Sum	43	6	6
	Mean	0.83	0.12	0.12
Friday	Sum	36	12	13
	Mean	0.69	0.23	0.25
Saturday	Sum	36	4	30
	Mean	0.69	0.08	0.58
Sunday	Sum	42	6	32
	Mean	0.79	0.11	0.60
Total	Sum	276	47	92
	Mean	0.76	0.13	0.25

Notes: This table presents descriptive statistics by day of week for the number of anti-Muslim hate crimes, the number of Trump's tweets about Muslims and the number of Trump's golf outing for the sample used in the instrumental variable regressions (2017 only).

Table 2.33: Robustness Time Series Regressions

	Baseline (1)	Add 7 lagged dependent variables (2)	Add golf holiday control (3)	Add previous week golf control (4)	Use Trump Tweet dummy (5)	Use only NYT golf count (6)	Use alternative golf count (7)
Panel A: OLS - Log(Hate crimes against Muslims) in t+2							
Log(Muslim Trump tweets)	0.130*	0.148**	0.128*	0.127*	0.106	0.130*	0.130*
	(0.069)	(0.069)	(0.069)	(0.069)	(0.074)	(0.069)	(0.069)
Panel B: First Stage - Log(Trump tweets about Muslims)							
Trump golfs	0.102***	0.098***	0.129***	0.094***	0.118***	0.095***	0.098***
	(0.027)	(0.027)	(0.031)	(0.027)	(0.033)	(0.028)	(0.027)
Panel C: Reduced form - Log(Hate crimes against Muslims) in t+2							
Trump golfs	0.165**	0.164**	0.163**	0.163**	0.165**	0.168**	0.155**
	(0.071)	(0.080)	(0.078)	(0.072)	(0.071)	(0.068)	(0.071)
Panel D: 2SLS - Log(Hate crimes against Muslims) in t+2							
Log(Muslim Trump tweets)	1.617**	1.682*	1.269**	1.631**	1.398*	1.764**	1.571*
	(0.779)	(0.935)	(0.633)	(0.821)	(0.716)	(0.824)	(0.809)
Weak IV 95% AR confidence set	[0.31; 4.01]	[0.29; 4.55]	[0.20; 2.96]	[0.27; 4.9]	[0.34; 3.74]	[0.54; 4.62]	[0.21; 4.05]
Fixed effects (month, day of week)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trends	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	363	358	363	363	363	363	363
R^2	0.213	0.209	0.339	0.207	0.204	0.149	0.231
Robust F-stat.	13.15	12.24	15.95	13.10	11.76	10.61	12.61

Notes: This table presents OLS and IV regressions where the dependent variable is the number of hate crimes against Muslims on any given day based on FBI data. We use a dummy for days on which President Donald Trump golfs used as an instrument for his tweets about Muslims. Column 2 controls for seven lags of the dependent variable. Column 3 controls for golf days that are part of a golf "holiday", which we define as Trump golfing for more than three consecutive days. Column 4 controls for whether Trump golfed in the previous week. Column 5 replaces the number of Muslim Trump tweets with a dummy for whether Trump sends any tweet about Muslims. Column 6 replaces the main measure $Trump\ golfs$ with one that only uses information from the New York Times (ignoring that contained in his presidential schedule). Column 7 uses an alternative golf count that incorporates information from *trumpgolfcount.com*. Column 8 presents an alternative specification where we cluster standard errors by week (ignoring serial correlation). The sample year is 2017, for which we have information on Trump's golfing. All regressions include day-of-week and year-month dummies, linear and quadratic time trends as well as a dummy for whether Trump's golfing is the first of a series of golf days. See online appendix for more details on data and variable construction. Newey-West standard errors are reported in parentheses except in column 8. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) with the Stata package from Sun (2018). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.34: Time Series - Split by Type of Hate Crime

	Any (1)	Vandalism (2)	Theft (3)	Burglary (4)	Robbery (5)	Assault (6)
Panel A: OLS - Log(Hate crimes against Muslims) in t+2						
Log(Muslim Trump tweets)	0.130*	0.023	0.023	0.093**	0.011	0.033
	(0.069)	(0.053)	(0.033)	(0.042)	(0.014)	(0.061)
Panel B: Reduced form - Log(Hate crimes against Muslims) in t+2						
Trump golfs	0.165**	0.139**	-0.003	0.022	-0.007	0.075
	(0.071)	(0.057)	(0.014)	(0.016)	(0.013)	(0.069)
Panel C: 2SLS - Log(Hate crimes against Muslims) in t+2						
Log(Muslim Trump tweets)	1.617**	1.363**	-0.033	0.216	-0.065	0.741
	(0.779)	(0.629)	(0.132)	(0.148)	(0.131)	(0.692)
Weak IV 95% AR confidence set	[0.31; 4.01]	[0.30; 3.29]	[-0.31; 0.27]	[-0.09; 0.58]	[-0.44; 0.16]	[-0.56; 2.59]
Fixed effects (month, day of week)	Yes	Yes	Yes	Yes	Yes	Yes
Time trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	363	363	363	363	363	363
R^2	0.213	-0.697	0.032	-0.026	0.004	0.291
Robust F-stat.	13.15	13.15	13.15	13.15	13.15	13.15

Notes: This table presents OLS and IV regressions where the dependent variable is the number of hate crimes against Muslims on any given day based on FBI data. We use a dummy for days on which President Donald Trump golfs used as an instrument for his tweets about Muslims. The sample year is 2017, for which we have information on Trump's golfing. All regressions include day-of-week and year-month dummies, linear and quadratic time trends as well as a dummy for whether Trump's golfing is the first of a series of golf days. See online appendix for more details on data and variable construction. Newey-West standard errors are reported in parentheses. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) with the Stata package from Sun (2018). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.35: Robustness Time Series Regressions - Timing of Effect

	t-1	t	t+1	**Baseline t+2**	t+3	t+4	t+5	t+6	t+7
	(1)	(2)	(3)	**(4)**	(5)	(6)	(7)	(8)	(9)
Panel A: OLS - Log(Hate crimes against Muslims)									
Log(Muslim Trump tweets)	0.112	0.008	0.084	**0.192****	-0.126*	-0.036	-0.162*	-0.047	0.030
	(0.100)	(0.111)	(0.102)	**(0.077)**	(0.075)	(0.100)	(0.085)	(0.081)	(0.093)
Panel B: Reduced form - Log(Hate crimes against Muslims)									
Trump golfs	0.079	0.048	0.077	**0.165****	0.097	0.085	-0.022	0.149**	0.144***
	(0.064)	(0.071)	(0.074)	**(0.071)**	(0.081)	(0.073)	(0.066)	(0.065)	(0.054)
Panel C: 2SLS - Log(Hate crimes against Muslims)									
Log(Muslim Trump tweets)	0.774	0.472	0.759	**1.617****	1.059	0.912	-0.224	1.500**	1.450**
	(0.642)	(0.648)	(0.725)	**(0.779)**	(0.854)	(0.729)	(0.682)	(0.729)	(0.692)
Weak IV 95% AR confidence set	[-0.43; 2.49]	[-1.13; 1.69]	[-0.60; 2.55]	**[0.31; 4.01]**	[-0.72; 3.34]	[-0.75; 2.72]	[-2.18; 1.06]	[0.27; 3.88]	[0.42; 3.99]
Fixed effects (month, day of week)	Yes	Yes	Yes	**Yes**	Yes	Yes	Yes	Yes	Yes
Time trends	Yes	Yes	Yes	**Yes**	Yes	Yes	Yes	Yes	Yes
Observations	363	364	364	**363**	362	361	360	359	358
R^2	0.430	0.482	0.469	**0.213**	0.324	0.399	0.519	0.181	0.222
Robust F-stat.	13.08	13.02	13.02	**13.15**	9.467	9.876	10.28	10.65	10.62

Notes: This table presents OLS and IV regressions where the dependent variable is the number of hate crimes against Muslims on any given day based on FBI data. Each column presents the results from a different regression, where the dependent variable is defined for the period in the top column. Column 4 is equivalent with column 6 in Table 2.6. We use a dummy for days on which Trump golfs used as an instrument for his tweets about Muslims. The sample year is 2017, for which we have information on Trump's golfing. All regressions include day-of-week and year-month dummies, linear and quadratic time trends, dummies for terror attacks in the US, Europe or the rest of the world, as well as a dummy for whether Trump's golfing is the first of a series of golf days. See online appendix for more details on data and variable construction. Newey-West standard errors are reported in parentheses. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) with the Stata package from Sun (2018). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.36: Robustness Controls

	Baseline (1)	Lagged dependent variable (2)	Federal holiday control (3)	Google search control (4)	Terror attack control (5)	Total tweets control (6)
Panel A: OLS - Log(Total number of Muslim TV mentions) in t+1						
Log(Muslim Trump tweets)	0.700***	0.301***	0.702***	0.631***	0.575***	0.700***
	(0.095)	(0.073)	(0.095)	(0.088)	(0.100)	(0.092)
Panel B: Reduced form - Log(Total number of Muslim TV mentions) in t+1						
Trump golfs	0.299**	0.142**	0.296**	0.311**	0.278**	0.297**
	(0.131)	(0.070)	(0.131)	(0.123)	(0.119)	(0.128)
Panel C: 2SLS - Log(Total number of Muslim TV mentions) in t+1						
Log(Muslim Trump tweets)	2.958***	2.108*	2.869***	3.028***	3.276**	3.042***
	(1.014)	(1.136)	(0.995)	(0.941)	(1.433)	(1.082)
Weak IV 95% AR confidence set	[0.85; 5.27]	[0.20; 6.27]	[0.80; 5.13]	[1.07; 5.36]	[0.91; 7.70]	[0.79; 5.72]
Fixed effects (month, day of week)	Yes	Yes	Yes	Yes	Yes	Yes
Time trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	364	363	364	364	364	364
R^2	0.961	0.976	0.963	0.960	0.956	0.960
Robust F-stat.	13.02	8.928	13.39	13.39	10.77	12.05

Notes: This table presents OLS and IV regressions where the dependent variable is the number of times Muslims are mentioned on TV on a given day. We use a dummy for days on which Trump golfs used as an instrument for his tweets about Muslims. The results for *Total Coverage* shown here are based on Fox News, CNN, and MSNBC. The results for the individual channels are available upon request. Column 2 controls for one lag of the dependent variable and column 3 for a dummy that tags federal holidays. Column 4 controls for the first principal component of Google searches for Islam-related terms. Column 5 controls for the number of terror attacks in the US, Europe, or other countries. Column 6 controls for the total number of tweets by Trump. The sample year is 2017, for which we have information on Trump's golfing. All regressions include day-of-week and year-month dummies, linear and quadratic time trends as well as a dummy for whether Trump's golfing is the first of a series of golf days. See online appendix for more details on data and variable construction. Newey-West standard errors are reported in parentheses. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) with the Stata package from Sun (2018). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.37: Summary Statistics for Time Series – Split at Campaign Announcement

	Before campaign announcement						After campaign announcement					
	Mean	SD	p50	Min	Max	N	Mean	SD	p50	Min	Max	N
Trump tweets												
Muslim Trump tweets (1+log)	0.03	0.16	0.00	0.00	1.39	2234	0.10	0.30	0.00	0.00	1.79	930
Total Trump tweets (1+log)	1.56	1.37	1.61	0.00	5.00	2234	2.27	0.72	2.30	0.00	4.54	930
Muslim Trump tweets (dummy)	0.04	0.20	0.00	0.00	1.00	2234	0.11	0.31	0.00	0.00	1.00	930
Hate crimes (1 + natural logarithm)												
Muslims	0.26	0.39	0.00	0.00	1.61	2234	0.47	0.48	0.69	0.00	1.95	930
All hate crimes	2.84	0.31	2.89	1.10	3.61	2234	2.89	0.30	2.89	1.79	3.76	930
Ethnicity (incl. Hispanic)	0.51	0.48	0.69	0.00	2.30	2234	0.40	0.45	0.00	0.00	2.08	930
Race	2.13	0.40	2.20	0.00	3.14	2234	2.17	0.40	2.20	0.69	3.04	930
Sexual orientation	1.35	0.51	1.39	0.00	2.56	2234	1.28	0.50	1.39	0.00	2.40	930
Religion (excl. Muslims)	1.22	0.55	1.39	0.00	2.71	2234	1.26	0.53	1.39	0.00	2.89	930
Other control variables												
Terror attack in the US	0.00	0.02	0.00	0.00	1.00	2234	0.01	0.07	0.00	0.00	1.00	930
Terror attack in Europe	0.00	0.04	0.00	0.00	1.00	2234	0.04	0.19	0.00	0.00	1.00	930
Terror attack elsewhere	0.02	0.14	0.00	0.00	3.00	2234	0.15	0.43	0.00	0.00	3.00	930

Notes: This table presents descriptive statistics for the OLS sample. The sample is split into the period before and after June 16, 2015 when Trump announced his presidential campaign. *1+log* or *1+natural logarithm* means that the logarithm of any variable is calculated with 1 added inside. The data on hate crimes come from the FBI hate crime statistics. The sources for the number of terror attacks is the Global Terrorism Database. See the online appendix for more details on data and variable construction.

Table 2.38: Time Series Regression Full Period

	Baseline (1)	Add lagged dependent variable (2)	Add terror attack control (3)	Add total tweets control (4)	Use Trump Tweet dummy (5)
Panel A: Before campaign announcement					
Log(Muslim Trump tweets)	0.017	0.018	0.019	0.015	0.053
	(0.018)	(0.018)	(0.018)	(0.019)	(0.098)
Observations	2,234	2,232	2,233	2,234	2,234
R^2	0.026	0.027	0.028	0.026	0.026
Panel B: After campaign announcement					
Log(Muslim Trump tweets)	0.108**	0.104***	0.090**	0.094**	0.307**
	(0.042)	(0.039)	(0.041)	(0.041)	(0.132)
Observations	930	928	929	930	930
R^2	0.079	0.082	0.092	0.082	0.077
Fixed effects (month, day of week)	Yes	Yes	Yes	Yes	Yes
Time trends	Yes	Yes	Yes	Yes	Yes

Notes: This table presents OLS regressions where the dependent variable is the number of hate crimes against the group in the top row on any given day based on FBI data. The sample is split into the period before and after June 16, 2015 when Trump announced his presidential campaign. All regressions include day-of-week and year-month dummies as well as linear and quadratic time trends. See online appendix for more details on data and variable construction. Newey-West standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.39: Time Series - Split by Motivating Bias

	All	Hispanic	Other Ethnicity	Race	Sexual Orientation	Religion (excl. Muslims)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: OLS - Log(Hate crimes)						
Log(Muslim Trump tweets), t-1	0.077*	0.007	0.247***	0.004	0.009	0.090
	(0.047)	(0.076)	(0.095)	(0.067)	(0.076)	(0.074)
Panel B: Reduced form - Log(Hate crimes)						
Trump golfs	0.013	-0.095	-0.017	0.043	0.024	0.011
	(0.045)	(0.084)	(0.084)	(0.065)	(0.070)	(0.069)
Panel C: 2SLS - Log(Hate crimes)						
Log(Muslim Trump tweets), t-1	0.105	-0.739	-0.129	0.333	0.183	0.089
	(0.330)	(0.619)	(0.636)	(0.482)	(0.530)	(0.516)
Weak IV 95% AR confidence set	[-0.71; 0.73]	[-2.15; 0.55]	[-1.70; 1.07]	[-0.67; 1.34]	[-0.92; 1.39]	[-1.19; 1.06]
Fixed effects (month, day of week)	Yes	Yes	Yes	Yes	Yes	Yes
Time trends	Yes	Yes	Yes	Yes	Yes	Yes
Observations	363	363	363	363	363	363
R^2	0.993	0.618	0.479	0.977	0.898	0.901
Robust F-stat.	15.95	15.95	15.95	15.95	15.95	15.95

Notes: This table presents OLS and IV regressions where the dependent variable is the number of hate crimes against the group in the top row on any given day based on FBI data. We use a dummy for days on which Trump golfs used as an instrument for his tweets about Muslims. The sample year is 2017, for which we have information on Trump's golfing. All regressions include day-of-week and year-month dummies, linear and quadratic time trends as well as dummies for whether Trump's golfing is the first of a series of golf days or part of a "golf holiday" (longer than three days). See online appendix for more details on data and variable construction. Newey-West standard errors are reported in parentheses. Weak IV 95% Anderson-Rubin (AR) confidence sets are calculated using the two-step approach of Andrews (2018) with the Stata package from Sun (2018). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.40: Time Series Regression Full Post-Campaign Period: Split by Motivating Bias

	All (1)	Muslim (2)	Ethnicity (3)	Race (4)	Sexual Orientation (5)	Religion (excl. Muslims) (6)
Panel A: Before campaign announcement						
Log(Muslim Trump tweets)	0.013	0.017	-0.001	0.005	-0.012	0.015
	(0.020)	(0.018)	(0.018)	(0.022)	(0.021)	(0.022)
Observations	2,234	2,234	2,234	2,234	2,234	2,234
R^2	0.232	0.026	0.016	0.153	0.107	0.064
Panel B: After campaign announcement						
Log(Muslim Trump tweets)	0.027	0.108**	-0.030	0.027	-0.006	-0.056
	(0.039)	(0.042)	(0.030)	(0.028)	(0.033)	(0.039)
Observations	930	930	930	930	930	930
R^2	0.196	0.079	0.034	0.155	0.077	0.119
Fixed effects (month, day of week)	Yes	Yes	Yes	Yes	Yes	Yes

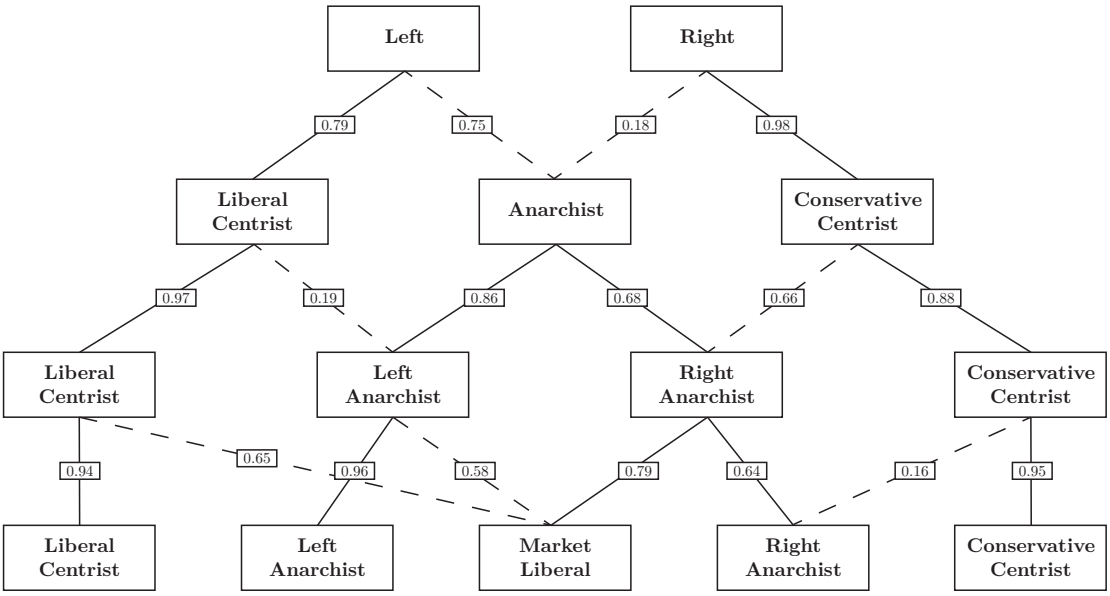
Notes: This table presents OLS regressions where the dependent variable is the number of hate crimes against the group in the top row on any given day based on FBI data. The sample is split into the period before and after June 16, 2015 when Trump announced his presidential campaign. All regressions include day-of-week and year-month dummies. See online appendix for more details on data and variable construction. Newey-West standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

2.9.5 Appendix 5: Additional Bartik Evidence

Table 2.41: Bartik Timing Results

	(1) OLS	(2) OLS	(3) Reduced Form	(4) Reduced Form
F4.Muslim Trump Tweet × Twitter Usage	-0.002	-0.002	-0.002	-0.002
	(0.003)	(0.003)	(0.002)	(0.002)
F3.Muslim Trump Tweet × Twitter Usage	-0.001	-0.001	-0.001	-0.001
	(0.004)	(0.004)	(0.004)	(0.004)
F2.Muslim Trump Tweet × Twitter Usage	0.003	0.002	0.004	0.004
	(0.003)	(0.003)	(0.003)	(0.003)
F.Muslim Trump Tweet × Twitter Usage	0.003	0.003	0.002	0.002
	(0.004)	(0.004)	(0.005)	(0.005)
Muslim Trump Tweet × Twitter Usage	0.003	0.004	0.007	0.007
	(0.004)	(0.004)	(0.004)	(0.004)
L.Muslim Trump Tweet × Twitter Usage	0.009**	0.010**	0.007*	0.008**
	(0.004)	(0.004)	(0.004)	(0.004)
L2.Muslim Trump Tweet × Twitter Usage	-0.000	0.001	0.002	0.002
	(0.004)	(0.004)	(0.003)	(0.003)
L3.Muslim Trump Tweet × Twitter Usage	0.002	0.003	0.001	0.001
	(0.002)	(0.002)	(0.003)	(0.003)
L4.Muslim Trump Tweet × Twitter Usage	-0.001	-0.000	-0.005	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)
County FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
County x Month FE	Yes	Yes	Yes	Yes
County X Day of Month FE	Yes	Yes	Yes	Yes
Pop. deciles x Day FE	Yes	Yes	Yes	Yes
7 lags dep. variable		Yes		Yes
Observations	2865576	2856252	2865576	2856252

Notes: This table presents OLS and reduced form regressions where the dependent variable is the log number of anti-Muslims hate crime in county c on day d. The independent variable is either the interaction Trump's anti-Muslim tweet with county-level Twitter usage or a reduced form/IV specification with our SXSW variables. The variables are standardized to have a mean of zero and standard deviation of one. All regressions include 4 leads and lags of Trump's anti-Muslim tweets. All regressions include population controls, county, day, county time month and county times day of month fixed effects. Later regression control also control for 7 lags of the dependent variable. Robust standard errors in parentheses are clustered by state. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
This table presents OLS and IV regressions where the dependent variable is We standardized the variables to have a mean of zero and standard deviation of one
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.



3) How Polarized are Citizens: Measuring Ideology from the Ground-up

Carlo Schwarz (University of Warwick)

Mirko Draca (University of Warwick)

3.1 Introduction

In political terms, we seem to be living in the midst of the proverbial 'interesting times'. Across established democracies there appear to be strong trends of political populism and ideological polarisation. In the US, a large body of evidence indicates that the political positions taken by elected representatives in legislatures have sharply polarised. For example, this is apparent in recent work examining partisanship in the use of political language (Jensen et al., 2012; Gentzkow et al., 2019b). In particular, Gentzkow et al. (2019b) isolate this increase as occurring from the mid-1990s onwards, a period when the nature of political communication changed as parties became more acutely strategic with their use of language. Further evidence of 'elite polarisation' is also found in the extensive literature (following Poole and Rosenthal, 1985) that has measured the evolving ideological positions of elected representatives using data on Congressional rollcall voting.

By comparison, the evidence about political polarisation amongst the general public (or 'citizens') is more contested than the findings that have emerged for political elites. In the US, contributions such as Fiorina and Abrams (2008) make the point that both the underlying distribution of views across issues and the level of self-identification with 'strong' political positions have been stable over time. Similar scepticism about citizen polarisation in the US is also evident in the studies of Glaeser and Ward (2006) and Ansolabehere et al. (2006), while a recent analysis by Kaplan et al. (2019) emphasises an important trend of rising within-state polarisation. In Europe, recent contributions by Algan et al. (2017) and Guiso et al. (2017) have documented a strong pattern of populist politics across the continent that appears to have roots in changing economic conditions. However, these populist trends are not necessarily symptomatic of ideological polarisation. For example, Algan et al. (2017) detect no significant shift in political positioning along the left-right scale in their cross-country sample and pick up a decline in close party identification.[63]

In this paper, we propose a new approach to measuring citizen ideology and political polarisation using unsupervised machine learning tools as applied to 'issue-position' data on individual political views. In short, the core of our approach is based on applying Latent Dirichlet Allocation (LDA) topic models (Blei et al., 2003) to individual-level survey responses across a typical range of social and economic issues. Topic models are mainly known in

[63]In terms of international comparisons, recent work by Boxell et al. (2020) finds that the US stands out internationally in terms of 'affective' polarisation (ie: dislike towards other political parties).

the social sciences for their use in the analysis of text data, in particular for their capacity in identifying the latent topic structure that underpins the generation of documents across various corpora. Applications of topic modelling have thus proliferated recently with empirical studies of text data across a range of social science questions (Gentzkow et al., 2019a). Within economics, the general approach we take here for analysing discrete, non-text data is closest to Bandiera et al. (2020)'s empirical model of behavioural manager 'types' in CEO time-use data.

Similarly, rather than analysing text we instead make individual-level responses from survey data the main objects of analysis, interpreting the latent topics as *political ideologies* that underpin the generation of individual political beliefs amongst the general public. The advantage of this particular approach is that it is based on a probabilistic generative model of ideology, allowing individual beliefs to be explained as mixtures of latent ideologies. As such, it is a concept of ideology that is directly empirical, that is, built up from the statistical pattern of political views across the population. Our approach also allows the identified citizen ideologies - defined practically as probability distributions over issue-positions - to evolve over time such that 'within' and 'between' shifts in ideology can be measured. In this way, we can move beyond ideologies defined based on political party affiliations.

We use this methodology to explore two main questions. Firstly, we ask: to what extent do the general public hold beliefs that can be summarised as statistically coherent 'ideologies'? Further to this point, to what extent do the latent ideologies found in the data conform to the traditional left-right ideological line that dominates both popular discourse and classic formal models in the spirit of Downs (1957)? This assumption of systematic coherence in political views within the population has been challenged by puzzles about citizen political views that have emerged from research on subjects such as preferences over redistribution (e.g Ashok et al., 2015)[64], as well as recent critiques of the principle of retrospective voting that have explored how people use different types of information in electoral decisions (Healy and Malhotra, 2013)[65].

The second main question we address is then: how do the empirically-based citizen ideologies we identify vary across countries and over time? In practical terms, this involves

[64]For example, data on evolving political views indicates that the demand for re-distribution via taxation is not increasing despite higher economic inequality. Ashok et al. (2015) show that, in US data, this cannot be explained by a general ideological shift to the right and focus their explanations on how re-distributive preferences vary by demographic sub-group.

[65]Achen and Bartels (2002) point out the apparent sensitivity of voters to arbitrary local events while contributions such as Wolfers (2002) and Leigh (2009) test for evidence of economic voting.

studying the factors that determine the ideological mixture of views held by citizens at the individual level and assessing the extent to which aggregate shifts can be explained in terms of changing demographics or other observables. Importantly, because our topic modelling approach allows for the mixed membership of individuals with respect to latent ideologies, it lets us parse individual ideological positions very finely. As a result, we can develop a measure of 'citizen slant' that captures the degree to which individuals weigh alternative ideologies within their overall beliefs. This measure captures, for example, the extent to which a given person is, say, 'a bit conservative and a bit liberal'. We then use this measure of slant to better characterise overall patterns of political polarisation. In particular, we put forward an analysis of multidimensional polarisation over more than two ideologies following the framework of Esteban and Ray (1994) and Duclos et al. (2004).

The main data source we use in our analysis is the cross-country World Values Survey (WVS) which provides a wide-ranging set of consistently asked questions from the late 1980s onwards. In answer to our first major question, a series of coherent citizen ideologies do indeed emerge from our modelling. A left-right dimension is strongly evident in the data but, alongside this, citizen confidence in institutions defines another major ideological dimension. We generically label the ideological types that are characterized by low confidence in institutions as 'anarchist' but note that the broad position that this type represents is consistent with the anti-establishment or populist positions that have been the focus of recent research (Acemoglu et al., 2013; Piketty, 2018; Rodrik). The anarchist label that we use is meant to avoid pejorative interpretations of terms such as populist[66] and emphasize opposition to current institutional structures as the defining feature of this ideological type.

Our unsupervised machine learning models allows us to document that ideological types emerge as a clear hierarchy of empirical ideologies as fit models with different numbers of types to the WVS. Inspired by the literature on topic cohesion (e.g. Chang et al., 2009), we propose a measure for the cohesion (quality) of ideologies based on a Normalized Pointwise Mutual Information (NPMI) criteria, as tested on hold-out samples of our WVS data. Based on this cohesion measure, our main empirical model of ideology takes the shape of a 4-type model. We label the 4 main types as Liberal Centrist, Conservative Centrist, Left Anarchist, and Right Anarchist.

Next, we use our findings regarding the structure of the ideologies to analyze the variation

[66]For example, see media critiques such as 'Populism: It's the BBC's new buzzword, being used to sneer at the 'uneducated' 17 million who voted for Brexit' from the UK's *Daily Mail* (Murray (2016)).

of ideologies across countries and time periods. Firstly, at the level of the latent ideologies, we find that our 4 main ideological types are stable over time with limited 'within ideology' changes, as measured by the weighting of different issue-positions. The most notable finding here is an increase in the intensity of socially liberal attitudes across most types. For example, the Conservative Centrist type shifts in their attitudes on issues such as homosexuality and abortion.

Secondly, we use the information on individual type shares (the mixture parameter in our LDA model) to measure how prevalent different ideologies are across countries and how this changes over time. The general pattern is consistent with the existing literature - for example, northern European countries are more liberal while countries with stronger religious traditions are more conservative. In turn, this is reinforced by a sensible pattern of correlations between individual-level characteristics and type shares (eg: women are more liberal and conservatism increases with age). Our main finding here is that the composition of the aggregate type shares is stable across time for most countries. However, a notable exception is the US where the total type share for the two Anarchist types increases from around 30% in the 1989-1993 wave to 50% by the fifth WVS wave in 2005-2009. The majority of this increase is accounted for by the Right Anarchist type.

The ideological type shares also have interesting relationships with variables representing self-positioning on the Left-Right scale and the probability of voting for 'populist' political parties. We find a strong relationship between type shares and Left-Right self-positioning - note here that the question on the positioning is excluded from the LDA model that defines the types. The ideological type shares also prove to be better predictors of populist voting than Left-Right self-positioning. For example, we estimate that an individual with a 50% type share in either the Left or Right Anarchist ideologies has a 38% higher probability of voting populist relative to the mean even after controlling for Left-Right self-positioning and other covariates.

The final part of our analysis then uses the outputs of the empirical LDA model to devise two further measures of ideological structure. The first 'citizen slant' measure we calculate provides a within-person measure of ideological concentration and is constructed following a basic Gini index logic. It directly exploits the mixed membership format of our unsupervised learning framework to capture how partisan individuals are in their ideological views. We find that the mean citizen slant across types, countries and years is relatively high at around 0.75 on a 0-1 scale. The degree of slant or within-person concentration has

also increased over the time window we consider. There is a slight increase in the case of Europe (of around 1% relative to the baseline in the initial wave) but much stronger shifts are apparent in the US. The rise in the US is also focused heavily on the Anarchist types (which increased their slant by around 15%) as well as the Centrist Conservative type (a 5% increase).

The second societal polarisation measure that we put forward builds on the framework of Esteban and Ray (1994) and Duclos et al. (2004). This framework allows us to develop a novel, multi-polar analysis of ideology in terms of own-group *identification* and between-group *alienation*. Practically, this is achieved by leveraging the information on relative group size within countries (where group membership is defined according to the dominant type share), alongside the other information from the LDA model outlined above. We find that changes in the level of polarisation over time are muted. Again, the US stands out as experiencing the sharpest increase, chiefly driven by the compositional change in type shares noted above. Interestingly, the nature of the US polarisation experience is more characteristic of a 'disappearing centre' driven by the growth of anarchist types than it is by a traditional left-right division.

Related Literature. The nature of this paper's main topic (pun unintended) means that it has connections with many literatures and contributions. Some areas to highlight are the following. Firstly, there is the literature on democratic politics and populism, with recent examples that include: Acemoglu et al. (2013), Algan et al. (2017), Buisseret and van Weelden (2017), Bursztyn et al. (c), Dal Bó et al., Dal Bó et al. (2018), Guiso et al. (2017), and Rodrik[67]. As discussed, our work sheds light on the potential long-run ideological underpinnings of these political trends in the population.

Secondly, there is fast-growing literature that studies aspects of ideology, policy-making and political communication using tools from machine learning and natural language processing. This includes the already noted Gentzkow et al. (2019b) and Jensen et al. (2012), as well as other text-based studies such as: Ash, Grimmer (2009), Hansen et al. (2014), Cagé et al. (2015) and Jelveh et al. (2015). Another branch of this overall literature (Blaydes and Grimmer (2013), Gross and Manrique-Vallier (2012), Wang et al. (2017), Munro and Ng (2019)) has also begun to explore the application of unsupervised learning tools to survey response data.

[67] A range of studies that have looked at the recent determinants of voting patterns are also relevant here: Becker et al. (2017), Dippel et al., Dorn et al. and Che et al..

Finally, there is a large literature that explicitly addresses polarisation and fractionalisation along political, ethnic and cultural lines. This literature often focuses on measuring group structure in societies and relating this to patterns of conflict. An indicative list includes: Alesina et al. (2003), Bossert et al. (2011), Caselli and Coleman (2013), Duclos et al. (2004), Esteban and Ray (1994) and Montalvo and Reynal-Querol (2005), Canen et al. (2020). Some recent work of interest here includes Bertrand and Kamenica, who measure 'cultural distance' between population sub-groups in the US and find a constant relationship amongst most outcomes and group splits. They do however note divergences in social attitudes based on political ideology and income. Desmet and Wacziarg (2018) also examine cultural distance, again finding stability across most dimensions.

Our paper extends the overall polarisation and fractionalisation literature by using a rigorous, data-driven approach for the definition and analysis of ideological groups. Recent works suggested that the left-right ideological model needs to be extended by either incorporating voters' identity (Gennaioli and Tabellini, 2019), a "globalists" vs "nativists" dimension (Piketty, 2018) or moral foundations Enke (2020). Our approach allows us to investigate which ideological dimensions and groups are most apparent in survey data across Western countries without reliance on predefined categories. Our results indicate that in addition to the left-right spectrum, trust in institutions appears to be a significant determining element of citizen ideology in the countries that we study.

The approach to identifying social sub-groups in a purely data-driven way has the potential to inform the emerging literature on identity politics (Atkin et al.; Grossman and Helpman; Shayo, 2009). Currently, this literature has focused on identity groups whose definition hinges on, ex-ante characteristics (eg: race, gender, income class). Our methodology shows that there is scope to define latent social sub-groups based on observable positions. We also note that the 'identification and alienation' framework used in our polarisation measures is directly analogous to key concepts in the identity literature (Akerlof and Kranton, 2000) and therefore provides metrics to study the potential frictions between politically social groups over time.

Structure. The paper is organized in the typical way. In section 2, we outline the main data used, namely the World Values Survey (WVS) as well as our approach to defining answers to survey questions as 'issue-positions'. Section 3 describes our unsupervised learning methodology for studying this issue-position data. This includes details on how we develop a

hierarchy of ideological types and select the optimal number of topics in our LDA models. Section 4 outlines the results and section 5 concludes.

3.2 Data

World Values Survey

For our main analysis, we use data from the World Values Survey (WVS) and the European Values Study (EVS). These surveys are an output of a global research project conducted by a large network of social scientists and run via a non-profit association based in Stockholm. They have been widely deployed in social science research and some prominent studies using the data include: Alesina et al. (2013, 2001); Blanchflower and Oswald (2008); Inglehart (1997); and Norris (2016).

The WVS consists of 6 waves from 101 countries while the EVS consists of 4 Waves from 48 countries. We construct what is formally known as the Integrated Value Survey (IVS) by combining the two datasets. The resulting dataset contains the 4 EVS waves and the corresponding waves 1, 2, 4 and 5 from the WVS[68]. For the sake of simplicity, we refer to this combination of the data as the 'World Values Survey (WVS)'.

The set of questions asked and countries covered differs across successive waves of the WVS. We, therefore, develop a sample of WVS observations based on the principle of capturing the widest range of consistently asked questions over waves and across countries[69]. Since the first wave has limited country and question coverage[70] we construct our sample from the second wave onwards and develop a set of 17 countries in Europe and North America (Austria, Belgium, Canada, Denmark, Finland, France, Germany, Iceland, Ireland, Italy, Malta, Netherlands, Portugal, Spain, Great Britain, United States, North Ireland) and 29 questions. The selected questions cover issues such as abortion, immigration, sexuality, the role of government, and confidence in institutions. The resulting dataset contains a total of 82,338 observations over 3 waves spanning the years from 1989 to 2010.

[68]There is no wave of EVS that corresponds to the 6th wave of the WVS. Therefore, our main dataset ends in 2010 to focus on a consistently defined set of repeated country-question cross-sections.

[69]We provide additional details on the selection of questions in Section 3.6.1. We also later find (in Section 3.6.5) that the basic structure of the ideological clusters we identify is robust to the inclusion or exclusion of questions.

[70]The countries Austria and Portugal, as well as 7 complete questions, are not contained in the first wave.

Construction of Features

As part of the data preparation, we unify the coding of the questions and convert them to the same scale. The intention here is to represent the answers to the survey questions as discrete 'features' for the subsequent topic modelling. Specifically, we recode the responses for each of the 29 questions into two indicator variables expressing either support or opposition to each issue, for example, an indicator variable if the person believes that abortion is justifiable and a second indicator variable if the person opposes abortion. In cases where a person expressed neither support or opposition to an issue both binary variables are coded as zero.

Summary statistics for the 58 recoded issue positions can be found in Table 3.1. Importantly, the features cover a broad range of salient political issues. Several questions deal with what would be typically classified as 'social issues' such as abortion, prostitution and attitudes towards minority groups while three questions deal with classic economic questions relating to the role of government, private sector competition and support for the welfare state. Finally, there is a set of questions dealing with confidence in a comprehensive set of social and political institutions.

The information in Table 3.1 indicates a rich mix of positions across political issues. There is a current of anti-foreigner sentiment with 12.3% of respondents preferring not to have immigrants as neighbours, and this is backed up by an overwhelming 60% endorsing a priority for native workers in the allocation of scarce jobs. However, most respondents either hold liberal or neutral views on leading social issues such as abortion and prostitution. There also is a widespread lack of confidence in key institutions, with only around 35-45% expressing a favourable view of the press, parliaments, the civil service and major companies.

3.3 Discovering Latent Ideology

We develop an approach based on machine learning topic models to investigate the pattern of responses in the WVS data in terms of a generating structure characterized by latent political ideologies. To be clear, we will define an 'ideology' as a probability distribution of issue-position responses across questions. Since the basic methods we use are most commonly applied to the analysis of text in terms of underlying topics or subjects, we first outline how we adapt the methods to study citizen ideology. We then describe an approach for model selection, that is, discerning the number of topics or ideological types that best describe the

Table 3.1: Summary Statistics, WVS Questions

Code	Question	Scale	Share For	Share Against
	On this list are various groups of people. Could you please mention any that you would not like to have as neighbors?	0,1		
A124_02	People of a different race		0.097	0.903
A124_06	Immigrants / foreign workers		0.123	0.877
A124_07	People who have AIDS		0.208	0.792
A124_08	Drug addicts		0.638	0.362
A124_09	Homosexuals		0.219	0.781
C002	Do you agree, disagree or neither agree nor disagree with the following statements?: "When jobs are scarce, employers should give priority to people of this country over immigrants."	1-3	0.600	0.305
E036	Rate your view on a 1 to 10 scale between the positions: "Private ownership of business and industry should be increased" vs. "Government ownership of business and industry should be increased"	1-10	0.506	0.255
E037	Rate your view on a 1 to 10 scale between the positions: "Government should take more responsibility to ensure that everyone is provided for" vs. "People should take more responsibility to provide for themselves"	1-10	0.376	0.469
E039	Rate your view on a 1 to 10 scale between the positions: "Competition is good. It stimulates people to work hard and develop new ideas" vs. "Competition is harmful. It brings out the worst in people"	1-10	0.613	0.215
	Could you tell me how much confidence you have in these organizations:	1-4		
E069_01	Church		0.519	0.481
E069_02	Armed Forces		0.567	0.433
E069_04	The Press		0.356	0.644
E069_05	Labour Unions		0.385	0.615
E069_06	The Police		0.704	0.296
E069_07	Parliament		0.413	0.587
E069_08	The Civil Services		0.451	0.549
E069_13	Major Companies		0.432	0.568
E069_17	Justice System / Courts		0.533	0.466
	Please tell me for each of the following actions whether you think it can always be justified, never be justified, or something in between:	1-10		
F114	Claiming government benefits		0.076	0.869
F115	Avoiding a fare on public transport		0.086	0.842
F116	Cheating on taxes		0.106	0.828
F117	Someone accepting a bribe		0.035	0.931
F118	Homosexuality		0.407	0.432
F119	Prostitution		0.196	0.663
F120	Abortion		0.348	0.458
F121	Divorce		0.496	0.280
F122	Euthanasia		0.418	0.430
F123	Suicide		0.149	0.730
G006	How proud are you of your nationality?	1-4	0.885	0.115

Notes: This table reports summary statistics for the recoded questions from the WVS. The third column reports the original coding of the question in the WVS. Questions with a binary or 1–4 coding are recoded into two indicator variables expressing either support or opposition to each issue. Questions with 1–3 or 1–10 allow for a neutral coding if the answer is coded as 3 or 5 in which case both indicator variables are coded as zero. The fourth (fifth) column contains the share of people are coded as a positive (negative) response to the question.

data. Finally, we discuss how we track changes over time within our overall topic model methodology.

3.3.1 Discovering Citizen Ideology via Latent Dirichlet Allocation (LDA)

The basis of our approach is Latent Dirichlet Allocation (LDA) (Blei et al., 2003) which can be summarised as a Bayesian hierarchical model that defines a probabilistic structure for joint distributions of observed data and latent generating factors. It was originally developed for the unsupervised classification of text data into a user-chosen number of topics.

In the context of text-based applications, LDA makes use of the fact that authors tend to use similar words when they talk about the same topic. For example, a text containing the words 'equilibrium' and 'preferences' is far more likely to be about economics than sports. LDA is therefore built on the principle of algorithmically classifying any corpus of text documents as a probabilistic mixture of underlying topics. Again, as an example, a document discussing a Pigovian tax might get classified as a mixture of a taxation topic and an environmental policy topic. Each LDA topic is defined as a probability distribution over words. A taxation topic, for example, might put high weights on the words 'tax', 'revenue' and 'IRS'.

Since the LDA algorithm itself does not provide any topic labels and the standard machine learning topic labelling approaches (e.g Lau et al., 2011; Aletras et al., 2014) are not applicable in our setting, it is up to the user to interpret and judge the focus of each topic. However, some metrics for 'topic coherence' are available for assessing the quality of a given topic model and to facilitate the choice of the optimal number of topics.

At its core, LDA is a clustering algorithm for discrete data. As a result, LDA can be used in non-text applications, for example, image classification tasks in the field of computer vision (e.g. Putthividhy et al., 2010)[71]. For our study, we apply LDA to the WVS survey responses of individuals. Instead of clustering frequently co-occurring words into a topic, LDA will combine issue positions that are frequently held together into an 'ideological type'. Each of the respondents in the WVS will also be classified as a mixture of ideological types based on their answers to questions, for example, as 20% 'conservative' and 80% 'liberal'.

Each ideological type will be described by a probability distribution over issue positions. This probability distribution describes how important the individual issue positions are for

[71] See also the collection by Airoldi et al. (2014) for a diverse set of applications of mixed membership modelling.

each ideological type. Our general approach is most closely related to Bandiera et al. (2020) who model CEO time use across discretely-defined activities.

The advantage of LDA in comparison to other clustering algorithms is that it provides a generative model of the data and thereby a quasi-microfoundation as each LDA parameter has a direct empirical interpretation. While both Principal Component Analysis (PCA) or Factor Analysis (FA) have been widely used to either identify the big 5 personality traits (e.g. Tupes and Christal, 1961; Norman, 1963) or the general intelligence factor g (e.g. Spearman, 1904), neither models the latent types of each individual directly. Moreover, both PCA and FA use linear transformations of the data while LDA allows for non-linear relationships. Overall, LDA is hence better suited for categorical data than either PCA or FA. Another advantage of LDA is that it is a mixed membership model which describes every observation as a mixture of types rather than in terms of some attachment to a single type or category, as in Latent Class Analysis, k-means, or spectral clustering.

Underlying our LDA model of citizen ideology is a probabilistic model which assumes that every individual $i \in I$ can be described as a probabilistic mixture of $t \in T$ topics or 'types'. These probabilities are contained in a vector θ_i of type proportions. The latent T types are described by 'type vectors' β_t with a question response profile for each of the Q questions. The entries in the type vector give the probability of holding a particular issue-position when drawing from a particular latent type. The generative process underlying the data is defined as:

1. For each individual i in the data draw ideological type proportions $\theta_i \sim Dir(\alpha)$, where α is a hyperpameter

2. For each of the $n \in N_i$ responses of individual i which we refer to as $r_{i,n}$:

   - Draw a type assignment $z_{i,n} \sim Mult(\theta_i)$
   - Draw a response $r_{i,n}$ from $P(r_{i,n}|z_{i,n}, \beta)$

Given this generative process the probability of the observed survey responses is:

$$\prod_{i=1}^{I} P(\theta_i|\alpha) \left(\prod_{n=1}^{N_i} \sum_{z_{i,n}} P(z_{i,n}|\theta_i) P(r_{i,n}|z_{i,n}, \beta) \right) \tag{3.9}$$

The first term describes how likely it is to observe an individuals ideological type proportions θ_i. The second term in brackets is the probability of observing the responses of individual's i. LDA identifies ideological types by finding parameter values for β and θ_i such that this probability is maximized. Due to dimensionality, simply maximizing this likelihood for the relevant parameters is computationally unfeasible. LDA therefore makes use of an approximate inference algorithm. We use the inference algorithm developed by Hoffman et al. (2010, 2013) and implemented by (Pedregosa et al., 2011).

In our application, the assumption of the independence of responses does not strictly hold. If a question has been answered the same question cannot be answered again by the same person. We discuss this in detail in Appendix 3.6.2, with specific reference to the survey data application of Gross and Manrique-Vallier (2012). In short, the inference of LDA is nonetheless still valid, since the bias in $P(r_{i,n}|z_{i,n})$ is identical for all types. Therefore, $z_{i,n}$ still represents the correct probability of a person belonging to one ideological type. Only the interpretation of the β vector changes. We provide an exposition of this specific point about the β vector in Appendix 3.6.3.

3.3.2 Determining the Optimal Number of Types

LDA makes it possible to estimate any number of ideological types. Therefore, the question of model selection is crucial for understanding which level of topic model best describes the data. In recent years, several methods for the understanding of topic cohesion in text data have been developed (e.g. Chang et al., 2009; Newman et al., 2010; Aletras and Stevenson, 2013; Lau et al., 2014). We modify these methods for the application to our "issue position" data. One advantage of our topic cohesion approach is that it is also applicable to any LDA model, especially non-text data. As such our approach could find application in any setting where previously the number of topics was simply chosen by the authors.

Our approach follows standard k-fold cross-validation principles. K-fold cross-validation works by fitting models to different parts or 'folds' of data. These models are then evaluated against each other based on an appropriate measure of model fit. As is standard in machine learning, the model with the best fit is used for analysis.

In our case, we first randomly split the data from the largest wave in our sample (wave 5) into 10 folds (each 10% of the data). Nine folds are then grouped into a training sample and the remaining becomes the test sample. Afterwards, we fit 10 LDA models with different

numbers of types (1 type up to 10 types) to the training sample. In each run of LDA, a different test sample is chosen and we evaluate the fit of each model relative to this hold-out data.

The optimal number of ideological types is then automatically chosen based on the cohesion of the generated types. A type is more cohesive if the issue positions with the largest weight for that type also frequently appear together in the held-out survey responses of WVS participants. The intuition behind this is that more cohesive ideological types should put more weight on issue positions that people frequently hold together, e.g. the co-occurrence of the views that abortion and suicide are not justifiable. This approach is preferable to evaluating the likelihood or the perplexity of the model in the hold out data, since the hold-out likelihood is not necessarily a good predictor for human judgment of topic cohesion (see for example Chang et al., 2009).

As a measure of co-occurrence of issue positions, we use Normalized Pointwise Mutual Information (NPMI). NPMI is defined as:

$$NPMI_{k,l} = \frac{PMI_{k,l}}{-\ln(p(k,l))} = \frac{\ln\left(\frac{p(k,l)}{p(k)\cdot p(l)}\right)}{-\ln(p(k,l))} \tag{3.10}$$

Pointwise Mutual Information (PMI) is simply defined as the log-ratio of the joint and marginal probabilities. Hence, PMI measures how probable it is that two features k and l appear together in comparison to how often we would expect them to appear together if the features were independent of each other. NPMI additionally normalizes PMI between $[-1, 1]$. If two features always appear together, their NPMI will be 1. In the case where two features never appear together, their NPMI will be -1.[72]

The average NPMI for all pairwise combinations of the B most important issue positions of an ideological type t is then given by:

$$\overline{NPMI_t} = \frac{\sum_k^B \sum_{l \neq k} (NPMI_{k,l})}{B \cdot (B-1)} \tag{3.11}$$

[72]More details on the topic cohesion literature and an example for the calculation of NPMI can be found in Section 3.6.4.

Similarly, the overall cohesion for a model with M ideological types can be calculated from the hold-out sample as:

$$Cohesion_m = \frac{\sum_t^M \overline{NPMI_t}}{M} \tag{3.12}$$

Follow the findings of Lau and Baldwin (2016) we average our measure of cohesion over different number of features $B \in (5, 10, 15, 20)$. As we discuss later, based on these scores we choose the 4 type LDA specification as our benchmark model, since it seems to best describe the pattern of responses across citizens.

Dynamic Type Models - Ideological Change Over Time.

The three waves of the WVS that we use stretch over 20 years. For our analysis, we want to allow for the ideological types to change over time. We do this by fitting LDA models separately to the 3 waves in our sample and only linking the ideological types together afterwards based on the similarity of their issue positions. Our approach is more generic than a dynamic topic model (Blei and Lafferty, 2006) or continuous topic model (Wang et al., 2008) since we neither impose any assumptions on the dynamics of the ideological types nor on the shares of the types over time. The general structure of our approach is most closely related to the topic chains suggested in Kim and Oh (2011) and has the advantage of allowing for completely different ideological patterns to emerge in each wave. But, as we will see, the ideological types in our WVS data displays a high degree of stability over time.

3.4 Results

We report our results across four linked sub-sections. In the first sub-section, we show the results of our LDA models in terms of different variants of type model - from 2-types to 5-types. The results here indicate a coherent hierarchy of types across the models such that types can be seen to 'split off' into related families as we move to higher-order models. The second sub-section then applies the NPMI model selection criteria outlined above to the different orders of type models with the conclusion that the 4-type model is the most preferred.

We then use the 4-type model as our main vehicle of analysis in the third sub-section, focusing on within-type and between-type differences over time. To guide the reader, this boils down to a close study of the β type vectors in the LDA model, that is, the probability

distribution of issue-positions per estimated type. In the final subsection, we focus on how the distribution of type proportions - essentially the θ_i values outputted by the LDA model - play out over countries and time. In turn, this leads to our analysis of within-person slant and country-level polarisation.

3.4.1 Hierarchy of Ideological Types.

In Table 3.2 we summarize the results of various orders of LDA models, reporting the 'top ten' features for each type. These top ten features represent the issue-positions with the highest probability values in the β type vectors and are effectively the defining features of each ideological type. We present the results as separate panels in the table per order of type model. [73]

Panel (a) shows the results for the basic 2-type model in the first column. These two types are distinguished by stances on social issues - for example, a liberal attitude towards minority groups (eg: reporting 'no problems' with neighbours who are homosexuals or immigrants) by one type and conservative positions on social issues such as abortion and prostitution by the other type. We, therefore, label these types in panel (a) generically as 'Left' and 'Right'. Across the 58 features, the β topic vectors for these types have a correlation of 0.39, indicating that they have some common positions.

The second column of panel (a) then reports the top features for the 3-type model. Two 'Left' and 'Right' types distinguished mainly by their positions on social issues such as sexuality, race and abortion are still apparent. However, the most striking result from this model is the nature of the third type. Rather than being a simple mixture of the basic Left-Right types of the earlier model the third type draws on a qualitatively different set of issue-positions for its top features. Specifically, the third type draws heavily on features that represent low confidence in major institutions such as parliament, the civil service, the press and major companies. We provide a more detailed discussion of the rationale for our type labels in the next sub-section but here we flag type 3 as an 'Anarchist' type to reflect this type's opposition to the current workings of major social institutions. In contrast, the main left and right types in the 3-type model report confidence in institutions across the majority

[73] The type hierarchy we show in Table 3.2 relates to wave 5 only. We focus on this wave as it represents the latest version or 'most recent evolution' across time of our basic types. However, as we show across multiple exercises (cohesion-based model selection, pooled wave model), the type structure is very stable over time. For completeness, we report the top ten features for the pooled wave model in Appendix Table 3.13.

Table 3.2: Hierarchy of Types (Top Ten Features)

(a) 2-4 Type Model

2 Type Model	3 Type Model	4 Type Model
Left	**Liberal Centrist**	**Liberal Centrist**
No problem Neighbours: Homosexuals	Confidence: Police	Confidence: Police
No problem Neighbours: People different race	Confidence: Justice System/Courts	No problem Neighbours: Homosexuals
No problem Neighbours: People AIDS	No problem Neighbours: Homosexuals	No problem Neighbours: People different race
No problem Neighbours: Immigrants/foreign workers	No problem Neighbours: People different race	Justifiable: divorce
Justifiable: divorce	No problem Neighbours: Immigrants/foreign workers	Proud of nationality
Not Justifiable: someone accepting a bribe	No problem Neighbours: People AIDS	No problem Neighbours: People AIDS
Proud of nationality	Proud of nationality	Not Justifiable: someone accepting a bribe
Justifiable: homosexuality	Not Justifiable: someone accepting a bribe	No problem Neighbours: Immigrants/foreign workers
Justifiable: euthanasia	Confidence: The Civil Services	Not Justifiable: claiming government benefits
Not Justifiable: claiming government benefits	Not Justifiable: cheating on taxes	Confidence: Justice System/Courts
Right	**Conservative Centrist**	**Conservative Centrist**
Not Justifiable: someone accepting a bribe	Not Justifiable: abortion	Confidence: Police
Proud of nationality	Not Justifiable: prostitution	Confidence: Churches
Not Justifiable: suicide	Not Justifiable: suicide	Confidence: Armed Forces
Not Justifiable: cheating on taxes	Proud of nationality	Not Justifiable: suicide
Not Justifiable: prostitution	Not Justifiable: someone accepting a bribe	Not Justifiable: prostitution
Not Justifiable: avoiding a fare on public transport	Not Justifiable: avoiding a fare on public transport	Not Justifiable: abortion
Not Justifiable: claiming government benefits	Not Justifiable: cheating on taxes	Proud of nationality
Not Justifiable: abortion	Not Justifiable: claiming government benefits	Confidence: Justice System/Courts
No problem Neighbours: People different race	Not Justifiable: homosexuality	Not Justifiable: someone accepting a bribe
Confidence: Police	Not Justifiable: euthanasia	Confidence: The Civil Services
	Anarchist	**Left Anarchist**
	No Confidence: Civil Services	No Confidence: Churches
	No Confidence: Parliament	Justifiable: divorce
	No Confidence: Churches	No problem Neighbours: Homosexuals
	No Confidence: Major Companies	No problem Neighbours: People AIDS
	No Confidence: Justice System/Courts	No problem Neighbours: People different race
	No Confidence: The Press	No problem Neighbours: Immigrants/foreign workers
	No problem Neighbours: Homosexuals	No Confidence: Parliament
	No problem Neighbours: People different race	Justifiable: homosexuality
	No problem Neighbours: People AIDS	No Confidence: Armed Forces
	No Confidence: Labour Unions	No Confidence: Major Companies
		Right Anarchist
		No Confidence: Parliament
		No Confidence: Civil Services
		No Confidence: Justice System/Courts
		No Confidence: The Press
		No Confidence: Labour Unions
		No Confidence: Major Companies
		Not Justifiable: someone accepting a bribe
		Not Justifiable: claiming government benefits
		Not Justifiable: avoiding a fare on public transport
		Not Justifiable: cheating on taxes

(b) 4-5 Type Model

4 Type Model	5 Type Model
Liberal Centrist	**Liberal Centrist**
Confidence: Police	Confidence: Police
No problem Neighbours: Homosexuals	Confidence: Justice System/Courts
No problem Neighbours: People different race	Confidence: The Civil Services
Justifiable: divorce	Justifiable: divorce
Proud of nationality	Confidence: Parliament
No problem Neighbours: People AIDS	Proud of nationality
Not Justifiable: someone accepting a bribe	No problem Neighbours: People different race
No problem Neighbours: Immigrants/foreign workers	Confidence: Armed Forces
Not Justifiable: claiming government benefits	Not Justifiable: someone accepting a bribe
Confidence: Justice System/Courts	No problem Neighbours: Homosexuals
Conservative Centrist	**Conservative Centrist**
Confidence: Police	Not Justifiable: abortion
Confidence: Churches	Not Justifiable: euthanasia
Confidence: Armed Forces	Not Justifiable: prostitution
Not Justifiable: suicide	Not Justifiable: suicide
Not Justifiable: prostitution	No problem Neighbours: People different race
Not Justifiable: abortion	Confidence: Churches
Proud of nationality	No problem Neighbours: Immigrants/foreign workers
Confidence: Justice System/Courts	Confidence: Police
Not Justifiable: someone accepting a bribe	Not Justifiable: someone accepting a bribe
Confidence: The Civil Services	Not Justifiable: cheating on taxes
Left Anarchist	**Left Anarchist**
No Confidence: Churches	No Confidence: Armed Forces
Justifiable: divorce	No Confidence: Churches
No problem Neighbours: Homosexuals	No Confidence: Parliament
No problem Neighbours: People AIDS	No Confidence: Major Companies
No problem Neighbours: People different race	No Confidence: Police
No problem Neighbours: Immigrants/foreign workers	No problem Neighbours: Homosexuals
No Confidence: Parliament	No Confidence: Civil Services
Justifiable: homosexuality	No Confidence: Justice System/Courts
No Confidence: Armed Forces	No problem Neighbours: People AIDS
No Confidence: Major Companies	No problem Neighbours: People different race
Right Anarchist	**Right Anarchist**
No Confidence: Parliament	Against Neighbours: People AIDS
No Confidence: Civil Services	Against Neighbours: Homosexuals
No Confidence: Justice System/Courts	Against Neighbours: Immigrants/foreign workers
No Confidence: The Press	Against Neighbours: Drug addicts
No Confidence: Labour Unions	If Jobs scarce: priority to (nation) people
No Confidence: Major Companies	Not Justifiable: homosexuality
Not Justifiable: someone accepting a bribe	No Confidence: Parliament
Not Justifiable: claiming government benefits	Against Neighbours: People different race
Not Justifiable: avoiding a fare on public transport	Not Justifiable: suicide
Not Justifiable: cheating on taxes	Proud of nationality
	Market Liberal
	No Confidence: Parliament
	No Confidence: Civil Services
	No Confidence: The Press
	No problem Neighbours: Homosexuals
	No problem Neighbours: People different race
	No problem Neighbours: People AIDS
	Not Justifiable: claiming government benefits
	Not Justifiable: someone accepting a bribe
	No Confidence: Labour Unions
	No problem Neighbours: Immigrants/foreign workers

Notes: This table reports the 10 most important features based on the β vectors for a n-type LDA model, where $n \in \{2, 3, 4, 5\}$.

of features in this category. We label these types as 'Liberal Centrist' and 'Conservative Centrist' to reflect their contrasting positions on social issues but common pattern of support for established political institutions.[74]

The top features for the 4-type model are reported in the third column of panel (a), Table 3.2. The type structure continues to evolve here. Most notably, two anarchist types now become apparent, again distinguished by contrasting views on social issues but similar positions in terms of confidence (or lack thereof) in institutions. These are labelled 'Left Anarchist' and 'Right Anarchist' to reflect this.[75] Intuitively, the top ten features reported in panel (c) suggest a splitting of the Anarchist type from the 3-type model has occurred.[76]

We can validate this by examining the cross-model correlations in the weights on issue-positions in the β type vectors. These correlations are useful for indicating how close the individual types in the 4-type model are to those in the lower order 3-type order. We report these in Figure 3.1. In line with the intuitive 'eyeballing' of the top features, the Left Anarchist and Right Anarchist types are most strongly correlated with the Anarchist type from the 3-type model, with correlation measures of 0.86 and 0.68 respectively. This splitting of the Anarchist type is reinforced by the continuity in the Liberal Centrist and Conservative Centrist types as we go from the 3-type to 4-type model. These two types can be tracked across the different hierarchies of type model, with correlations of 0.97 (Liberal Centrist) and 0.88 (Conservative Centrist) across the models.

The top features for a further 5-type model are reported in panel (b) of Table 3.2. Additional nuances in the types become evident here. The set of Liberal Centrist, Conservative Centrist and Left Anarchist types remain intact relative to the 4-type model but there appears to a splitting of the Right Anarchist type. Two variants of the Right Anarchist emerge. One variant still expresses a lack of confidence in institutions but appears to be liberal on social issues and is economically liberal in terms of attitudes towards unions and the claiming of government benefits. We label this type as 'Market Liberal'[77]. The other variant of the Right Anarchist is not socially liberal, with a string of conservative positions on minorities and

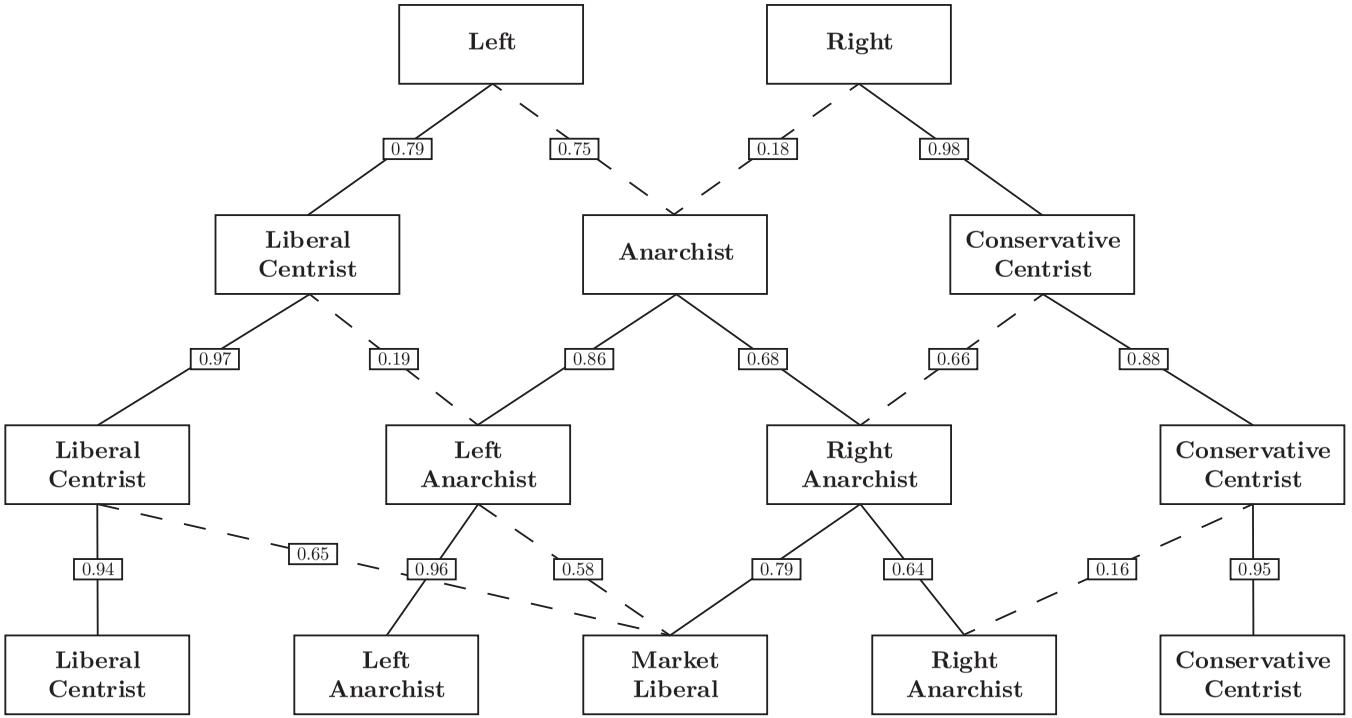
[74]Note here that the Conservative Centrist type in the 3-type model reports confidence in the churches, armed forces and police as its 11th-13th ranked features.

[75]The Right Anarchist type reports pride in his nationality, opposition to prostitution and drug addicts in the 11th-15th ranked features.

[76]Further evidence on which features separate the types for this 4-type model can be found in Table 3.16 in the Online Appendix, which reports the most important type differences.

[77]Our nickname for this anarchistic, pro-market and socially liberal type is 'George Mason University (GMU) Blogger'.

Figure 3.1: Hierarchy of Types as created by LDA



Notes: This graphic shows the hierarchy of types as created by Latent Dirichlet Allocation (LDA) for the different orders of topic model reported in Table 3.2.The values reported amongst the lines connecting the boxes record the correlation in the β issue-position probability vectors across types as a measure of type similarity.

social issues amongst its top ten features. The correlations indicate that this type is strongly correlated (0.64) with the original Right Anarchist from the 4-type model but negatively correlated with the Liberal Centrist (-0.195) and Left Anarchist types (-0.295) types.

Further results on potential 6-type and 7-type models are reported in Table 3.12 in Section 3.6.6. The basic set of types is preserved such that we can directly label the types in line with those identified in the 4-type and 5-type models. The evolution of the hierarchy is apparent in the further splitting of the Right Anarchist type (6-type model) and the emergence of a nihilistic 'Super Anarchist' type (7-type model).

Overall, the results presented above indicate that both the low-order (2 or 3 types) and higher-order (4 plus types) models offer plausible sets of types and, considered together, can be interpreted in terms of a coherent hierarchy. We next turn to the question of formal model selection using the NPMI framework outlined previously.

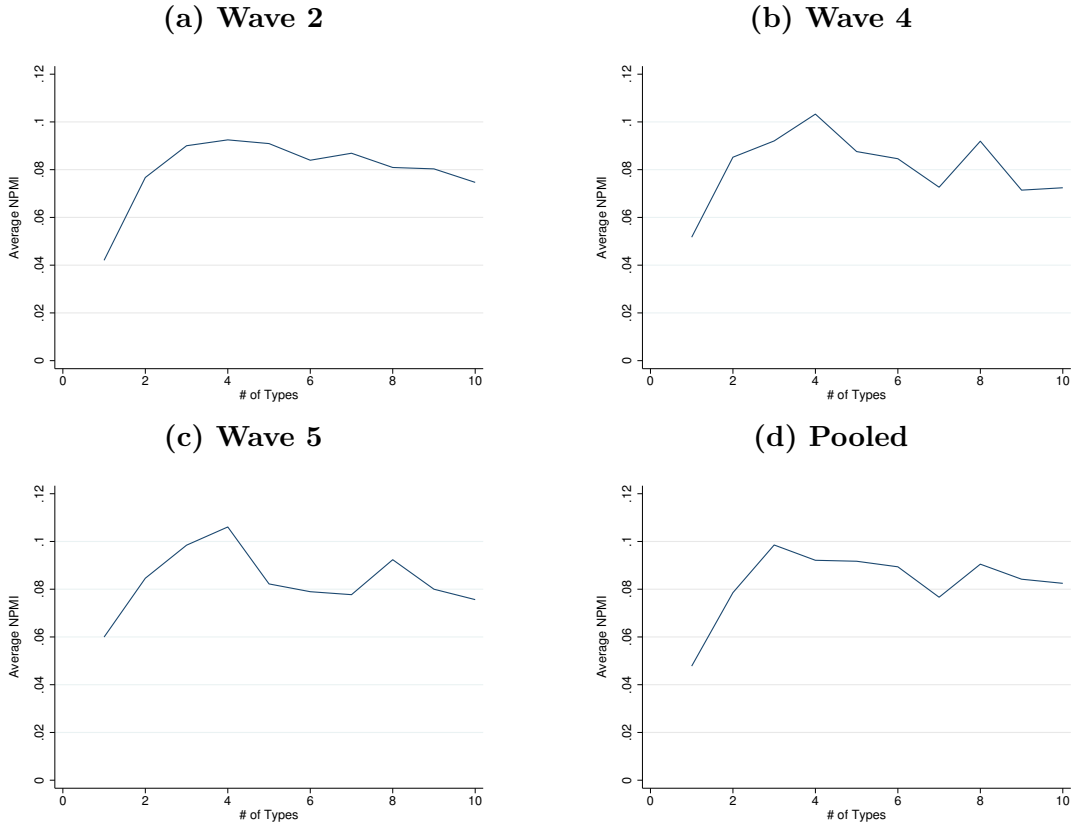
3.4.2 Model Selection and Type Labelling

Automatic Model Selection

Our NPMI framework for assessing model cohesion is based on comparing predictions of feature co-occurrence in hold-out data. Simply put, the approach asks: to what extent do the (say) top 10 features suggested by our type models occur together in data held out from the original estimation of the given model? Figure 3.2 reports the results of this exercise for all waves of the WVS. The x-axis denotes the order of model we are estimating (going from a 1-type model up to a 10-type model) while the y-axis denotes the resulting cohesion score.

In the 5th wave, the cohesion scores show an inverse u-shape pattern. At first, the cohesion score increases with the number of topics. After the number of topics increases beyond 4, the cohesion scores begin to fall. The exception being 8 topics for which we observe another increase of the cohesion score. Overall, the most cohesive models (M) appear to be either the 4 or 3 type models. We decided to use the 4 type model, since it delivers a higher cohesion score. In Figure 3.2 we also show that the 4 type model is preferred when we repeat the analysis for the other waves of the WVS. Only in the pooled model, there is a slight preference of the 3-type specification over the 4-type model. Given this evidence, our analysis from this point, therefore, employs the 4-type model composed of the Liberal Centrist, Conservative Centrist, Left Anarchist and Right Anarchist types.

Figure 3.2: Average Cohesion of Ideological Types for Different LDA Models



Notes: This figure show the topic cohesion scores calculated for models with $M \in \{1-10\}$ types for the 2nd/4th/5th wave as well as for a pooled model. The topic cohesion is calculated based on different numbers of features $B \in (5, 10, 15, 20)$. Afterwards, the average of the cohesion scores for different values of B is taken. See Section 3.4.2 for a more detailed description.

That said, note that our results do not hinge on this aspect of model specification and are qualitatively very similar if we use models of higher or lower numbers of types. The reason for this is that the types develop as part of a coherent hierarchy (see Figure 3.1). Moving to a 3 or 5 type model hence does not fundamentally change (for example) the prevalence of the Anarchist types in the data. We will return to this point about model specification again as we discuss various results (eg: on 'slant', polarisation, and links with populist voting).

Type Labelling

The labelling of our LDA-derived types is a point for discussion. An important advantage of our approach is that it is based on ideologies that emerge from the 'bottom-up' collection of views amongst the general public. The topics that we identify are empirical ideologies and may not necessarily have a tight mapping to traditional taxonomies of ideology[78].

Our labelling attempts to capture the main empirical differences in issue-positions between types. Note that we primarily use labels to simplify the exposition since we otherwise would have to refer to the types by numbers. Furthermore, the type labelling does not drive any of our results. Arguably, the main issue here is the labelling of types 3 and 4, which we have dubbed Left and Right Anarchist respectively. These two types are strongly distinguished by issue-positions that hinge on (low) confidence in institutions[79]. We use the term 'anarchist' to denote a pattern of opposition to current structures of political authority and hierarchical organization. That is, our use of the term is meant to be distinct from historical uses of the label, as per early socialist or syndicalist thinkers such as *inter alia*, Proudhon, Bakunin or Kropotkin.

Other plausible labels for these types are 'Populist' or 'anti-Establishment'. In particular, the fact that recent studies of populism (such as Algan et al. (2017)) have directly leveraged data on institutional trust provides some foundation for such a branding. However, we choose Anarchist as our label for this type because (i) it is a more generic and potentially neutral

[78]These taxonomies, covered in texts such as Vincent (2009) and Geoghegan (2003), are centred on 'classical' ideologies (eg: liberalism, conservatism, socialism) that are often explicitly articulated as bodies of thought by key writers (Locke, Burke, Stuart Mill, Marx), as well as modern 'post-materialist' ideologies (eg: environmentalism, feminism).

[79]See Appendix Table 3.16 for a breakdown of the largest differences in β issue-position weights across pairs of types. This shows the points of separation between the Centrist and Anarchist types.

term for the concept of an opposition to existing institutional structures[80], and (ii) the types that we identify are apparent from the early 1990s, thereby pre-dating the latest wave of populist politics. As a result, our empirical findings indicate that there may be some clear historical roots to the current populist trend, extending at least as far back as the late 1980s.

Alternative Models

In our appendix, we discuss two further modelling issues that relate to the quality of the information provided via the LDA framework. Firstly, in Appendix 3.6.5 , Table 3.10 we look at the sensitivity of our baseline 4-type model to the removal and addition of features. The basic model is very robust to the removal of features with types from iterative 'leave one out' models showing a high correlation with the types in the original model. The relative ordering of β weights is also preserved when we substantially widen the feature set (ie: add lots more questions - see Appendix 3.6.5 Table 3.11).[81] Both of these exercises provide reassurance that our overall baseline feature set is comprehensive enough to identify stable types in the data.

The second modelling issue that we examine (in Appendix 3.6.11) is a comparison of our LDA approach with other unsupervised learning methods. Specifically, we apply principal components analysis (PCA), factor analysis (FA) and k-means clustering to the same discretized feature data as our LDA models. As we discuss in Appendix 3.6.11, these alternative approaches are distinct from LDA in that they are linear methods and capture mixed membership relationships in a less explicit way. For example, a method such as PCA will pick out linear combinations of features with the highest degrees of variance in the data and therefore may not parse more complex data generating processes.

This is borne out in the types derived from these models which are reported in Appendix Tables 3.21, 3.22 and 3.23. The PCA models tend to identify conservative and anti-establishment types as part of the main model components, with no clear centrist or socially liberal types emerging. The FA and k-means models produce similar results. Further to this, no plausible hierarchy or 'family' of types emerges from these alternative models.

[80] As mentioned, the term 'populist' can be considered pejorative - see the blunt critique of UK's *Daily Mail* (Murray (2016)). The term 'anti-establishment' is subject to similar concerns, with competing claims of who the elite or establishment are. See, for example, Hume(2017) for a polemic about the liberal establishment and Jones(2014) for one about the conservative establishment.

[81] Among the questions in the widened feature set are many that no do not directly relate to political ideologies and which were therefore excluded from our baseline model. Further, some of the added questions are missing for close to 50% of the data and nearly a third of all questions are missing for more than 10% of the sample. Hence the model with the widened feature set requires extensive imputation and does not lend itself to be used as a baseline model.

Again, this provides reassurance that our LDA models - which are, after all, specifically intended for the analysis of discrete multinomial data - identify stable and interpretable types that are difficult to pin down using other methods.

Cross-Check Exercise Using the European Social Study

As further validation of our findings, we cross-check our results using the European Social Survey (ESS). This exercise serves two purposes. First, we want to demonstrate that our LDA methodology extends to other survey data. Second, we aim to understand if it is possible to recover similar ideological types using a different data set and a comparable set of questions. If our main LDA exercise is picking up valid latent types from the WVS data then this 'signal' should be apparent in other datasets.

Section 3.6.7 provides additional details on the ESS data and reports the full list of questions we selected for the exercise. Overall, the results from this exercise (see Section 3.6.7 for details) are striking. Although we use a completely different data set and varied the set of questions, the types that emerge from the ESS are broadly similar to those we identify in the WVS. In particular, we again find a split of the ideological types along the left-right spectrum and see a set of types characterized by their distrust in institutions.

3.4.3 Changing Ideologies?

Given the baseline 4-type model established above our next exercise examines within and between-type shifts across the different waves of the WVS. Our approach here is to estimate the 4-type model separately for each wave and compare the β type vectors over time.

The first point to note is that our main types are stable and repeat themselves across waves. This is evident in Table 3.3a which reports the correlations between the separately estimated types across waves. It is straightforward to pin down similar types across waves since the correlations are high, for example, the Liberal Centrist type showing a correlation of 0.97 between waves 2 and 4 or the Right Anarchist type reporting a correlation of 0.94 between waves 2 and 5.

These high correlations also imply that there are fairly limited 'within-type' shifts over time, as measured by the probability weights in the β type vectors. Since we are using the same issue-position features across waves we can directly report the shifts in probabilities

Table 3.3: Type Correlations

(a) Between-Wave Type Correlations

	Centrist Liberal Wave 2	Centrist Conservative Wave 2	Left Anarchist Wave 2	Right Anarchist Wave 2
Wave 4	0.973	0.985	0.963	0.981
Wave 5	0.935	0.963	0.943	0.939

Notes: This table reports the correlation of the β issue-position probability weights across types estimated in separate waves. That is, we identify 4 types in the initial Wave 2 (1989-1993) and correlate their β weights with the most similar types estimated separately on Waves 4 (1989-1993) and 5 (2004-2009).

(b) Within-Wave Type Correlations

Wave 2

	Liberal Centrist	Conservative Centrist	Left Anarchist	Right Anarchist
Liberal Centrist	1.000			
Conservative Centrist	0.418	1.000		
Left Anarchist	0.225	-0.525	1.000	
Right Anarchist	0.191	0.267	0.130	1.000

Wave 4

	Liberal Centrist	Conservative Centrist	Left Anarchist	Right Anarchist
Liberal Centrist	1.000			
Conservative Centrist	0.468	1.000		
Left Anarchist	0.322	-0.505	1.000	
Right Anarchist	0.251	0.289	0.178	1.000

Wave 5

	Liberal Centrist	Conservative Centrist	Left Anarchist	Right Anarchist
Liberal Centrist	1.000			
Conservative Centrist	0.523	1.000		
Left Anarchist	0.257	-0.408	1.000	
Right Anarchist	0.224	0.287	0.265	1.000

Notes: This table shows the correlation of the β issue-position probability weights amongst types in the same wave. That is, we estimate our 4 types using data on a single wave and then correlate the β weights across pairs of types in the same wave.

per feature. To facilitate the interpretation we have re-scaled the β vectors as described in Section 3.6.3[82].

These probabilities can be interpreted as statistics for the approximate 'likelihood of expression' for a given issue-position conditional on drawing on a latent type. For example, a (re-scaled) β weight of 0.46 for 'Confidence: Labor Unions' within the Liberal Centrist type indicates that an individual drawing on this type to form their issue-position will express confidence in this institution 46% of the time.

The ten largest shifts in probabilities for the Wave 2-5 difference are shown in Figure 3.3 for each type. The baseline numbers are also reported in Appendix Table 3.15 along with the

[82]Since the rescaling of the β vectors is a non-linear transformation, the changes between the re-scaled β vectors and unscaled β vectors can differ. See Section 3.6.3 for further technical details.

changes.[83] The most noticeable trend is an increase in socially liberal attitudes across types with, for example, the Conservative Centrist increasing their weights on issue-positions such as 'No problem neighbours: Homosexual' and 'No problem neighbours: People with AIDS'. Also notable is the Right Anarchist type, which shows higher confidence in the police and armed forces over time, along with more intense hostility towards immigration. Some of these changes are nominally large, with 10-15% increases in liberal attitudes on gay rights for the Conservative Centrist and 20-30% increases in confidence for the armed forces and police for the Right Anarchist. However, the overall changes in the β weights have not been pervasive enough to drastically shift the between-wave correlations evident in Table 3.3a[84].

The between-type differences can also be summarized using correlations across the β type vectors within the WVS waves and we show these in Table 3.3b. The increase in the intensity of socially liberal issue-positions is now most clearly seen via the increasing closeness between the Conservative Centrist type and the two left-wing types. Between waves 2 and 5 the negative correlation with the Left Anarchist type moderates (going from -0.525 to -0.408) while the correlation with the Liberal Centrist type strengthens (from 0.418 to 0.523). Hence, at the between-type level defined by the β-vectors, we can say that there has been some convergence in the overall ideologies driven in part by attitudes on social issues. Despite this convergence, note that the types remain clearly distinct and opposite to each other on many issues. As an illustration, in Appendix Table 3.16 we report the most important differences between the 4 types for the 5th wave.

3.4.4 Analysis of Type Shares

Correlates of Type Shares and Country Differences

We start the analysis of the θ_i individual type shares by studying the micro-level correlates. In particular, we estimate regressions of the individual type shares of the following form:

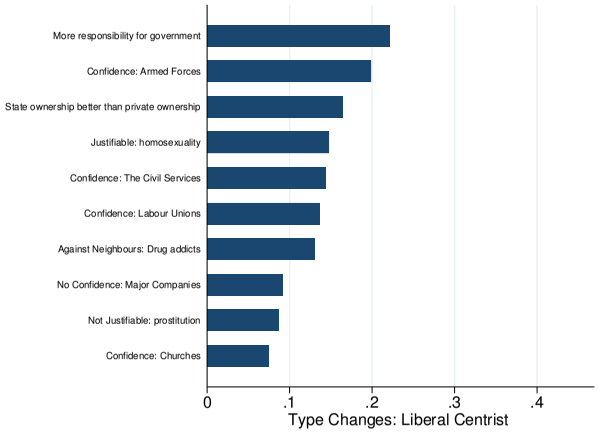
$$y^t_{icw} = X^{'}_{icw}\delta + \tau_w + \mu_c + \epsilon_{icw} \tag{3.13}$$

[83]Note that given the coding of each question into two features, the issues of decreasing importance will be approximately the opposite of the increasing features. Again, this relates to our adaptation of the LDA model for studying survey questions, which we outline in Section 3.6.2 and Section 3.6.3

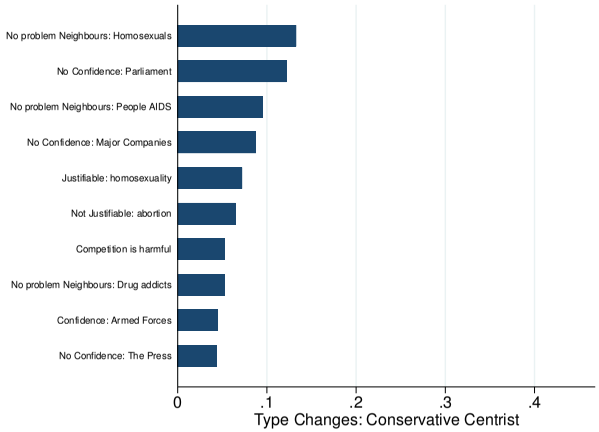
[84]In the case of Right Anarchist attitudes towards the police and armed forces it should be noted that this shift brings this type closer to the mean β for these issue-positions displayed by the two Centrist types.

Figure 3.3: Within-Type Changes in Issue-Position Weights (Wave2 to 5)

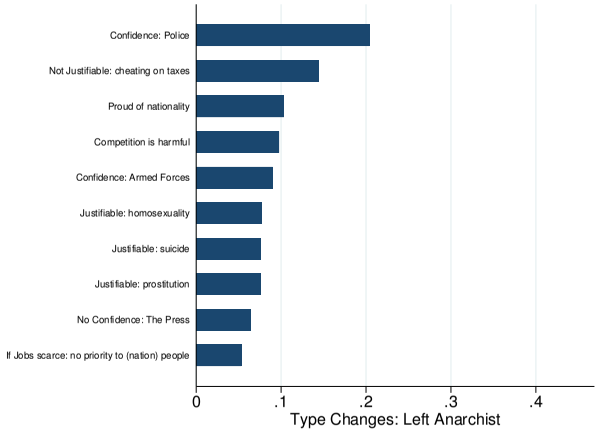
(a) Centrist Liberal



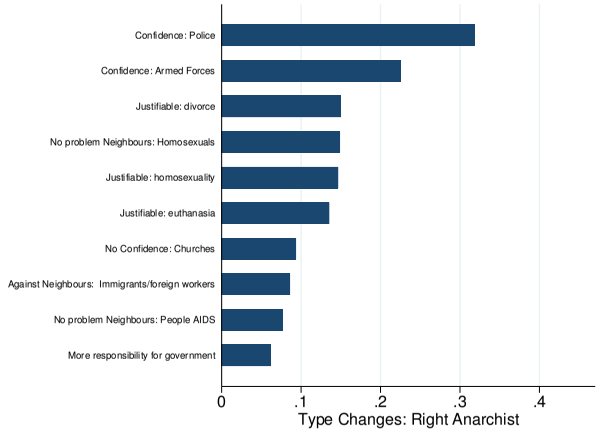
(b) Centrist Conservative



(c) Left Anarchist



(d) Right Anarchist



Notes: This figure reports the largest changes in the β issue position weights per ideological type from Waves 2 (1989-1993) to Wave 5 (2005-2009). We report the top ten changes per type amongst the 58 features. The scale is set to 0-0.4 to facilitate direct comparisons across types.

, where X_{icw} is a vector of covariates including the gender, age and the employment status.[85] τ_w and μ_c are wave and country fixed effects. The dependent variable y^t_{icw} is the share of type t of individual i in country c and wave w. Since the dependent variable is a continuous share the regression tells us how the intensity of ideological positions changes with different covariates. The results are reported in Table 3.4.

Table 3.4: Correlates of Individual-level Type Shares

	(1) Liberal Centrist	(2) Conservative Centrist	(3) Left Anarchist	(4) Right Anarchist
Female	0.015***	0.010***	-0.011***	-0.013***
	(0.002)	(0.002)	(0.002)	(0.002)
Age	-0.003***	0.004***	-0.003***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Unemployed	-0.073***	-0.003	0.049***	0.027***
	(0.005)	(0.005)	(0.004)	(0.005)
Wave 4	0.086***	-0.055***	0.007***	-0.039***
	(0.003)	(0.003)	(0.002)	(0.003)
Wave 5	0.069***	-0.076***	0.032***	-0.024***
	(0.003)	(0.003)	(0.002)	(0.003)
Observations	81,141	81,141	81,141	81,141
R-squared	0.143	0.105	0.111	0.055
Country FE	Yes	Yes	Yes	Yes

Notes: Each column reports the regression results for individual level regression of Equation (3.13). The dependent variable are the type shares for one of the 4 types created by LDA. Robust standard errors are used. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1. The data come from the World Value Survey and the European Value Survey.

We run four regressions corresponding to each type. These indicate some intuitively plausible correlations - women are more liberal and centrist, with a magnitude of 1.5% points, and the unemployed have higher shares in the two anarchist ideologies. Furthermore, there are clear shifts in the distribution of type shares over time. After conditioning on covariates, it is evident that the Liberal Centrist share increases by around 6.9% points after Wave 2 with the Conservative Centrist share falling by a similar amount. Following a similar pattern, the Left Anarchist share rises in Waves 4 and 5 while the Right Anarchist share falls.

Hence, across the sample of countries, the net result is a growth in the share of the two left-wing types (ie: Liberal Centrist and Left Anarchist). However, there are also significant country-level factors evident from the individual-level analysis. The country fixed effects in Table 3.4 account for 50-75% of the explained variation and we plot the country-level means by type in Figure 3.4. This again shows some expected relationships - northern European

[85] We only use a limited number of covariates in this exercise because these are the most complete ones available in the WVS in terms of missing values. When we run similar exercises with additional variables on reduced samples (circa N = 50,000) we get similar results (eg: low incomes are positively correlated with Anarchist shares, high education is positive with Liberal Centrism). These results available on request.

countries (eg: Denmark, Finland, Netherlands) are more liberal while countries with strong religious traditions (Malta, Ireland, Portugal) are more conservative.

To summarize the changes across countries over time we implement some splits along different ideological dimensions. Firstly, in Figure 3.5a we examine the left-right distinction and pool the type shares for the left-wing Liberal Centrist and Left Anarchist types.[86] The plot of changes in these pooled type shares between Waves 2 (1989-1993) and 5 (2005-2009) shows that most countries have moved left ideologically, with a mean shift of 8% points. In Figure 3.5b we then plot the changes for the pooled Left and Right Anarchist types. This provides an indicator of the overall strength of anti-establishment ideological sentiment across countries. The results show a large increase in the Anarchist type shares for the US (around 16% points), with significant increases also evident for the Anglo-Celtic domains (Great Britain, Northern Ireland) and the Netherlands. In net terms, however, the anarchist trend is more muted across countries.

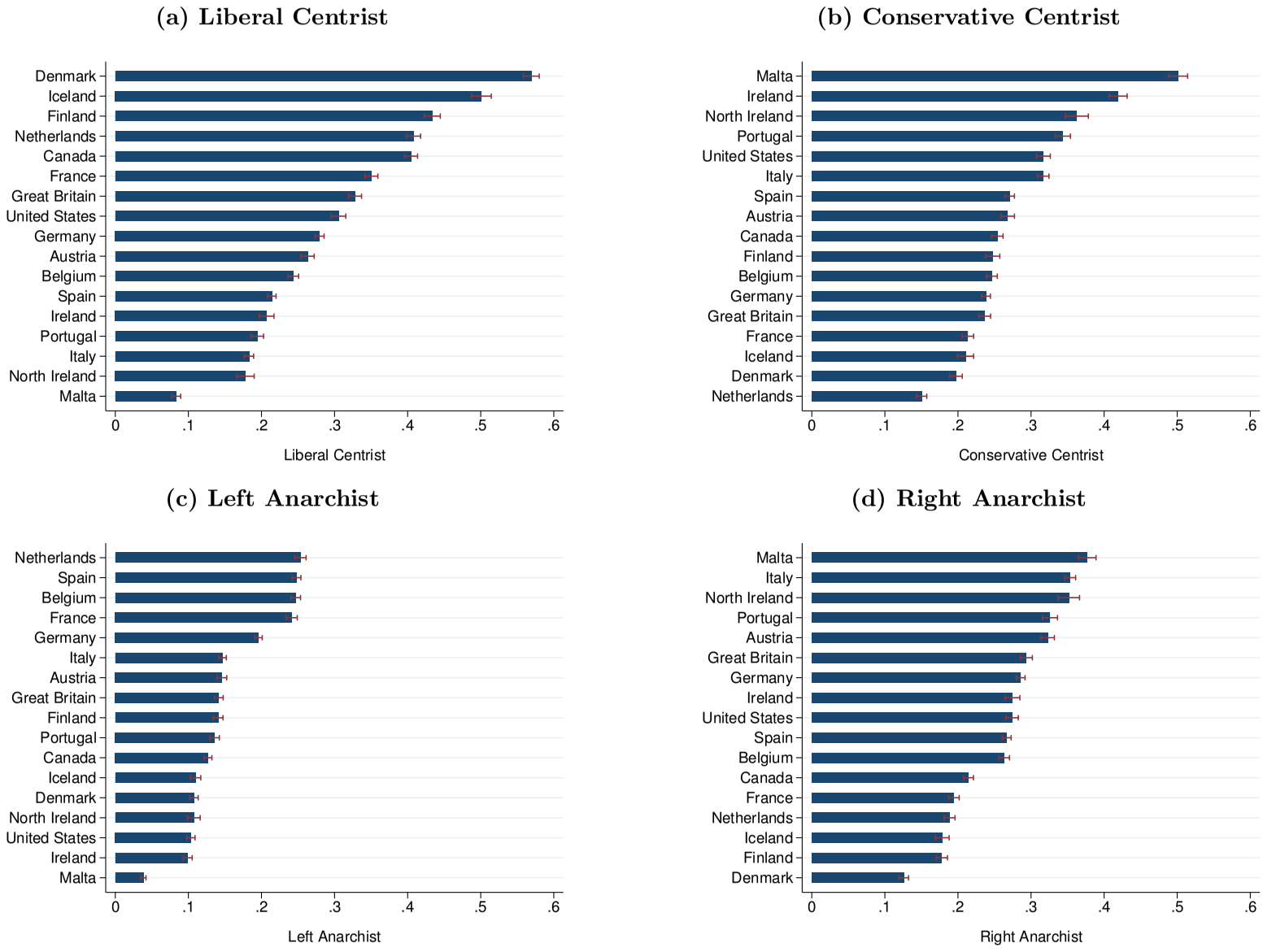
In Figure 3.6, we further probe the sharp increase in the Anarchist ideologies for the US. The clearest development is the growth in the US Right Anarchist share, which increases from a 24.9% share in Wave 2 (1989-1993) to 36.5% in Wave 5 (2005-2009). Note here however that this increase took place as a single step change between Waves 4 (1999-2004) and 5 (2005-2009). By comparison, the rise of the US Left Anarchist share from 7.9% to 12.7% was more gradual across the waves.

As discussed in Section 3.2, the set of countries available in wave 6 (post 2010) of the WVS is limited. However, in Section 3.6.8 we provide evidence that for the available countries the patterns in the 6th wave are consistent with the trends we document for wave 5. For example, the Anarchist type shares in the US remain at a high level.

Overall, these country-level findings are generally consistent with other international studies of shifts in political attitudes (Inglehart (1997); Inglehart et al. (2010)). Taken together with the within-type analysis, the basic message on the ideological change that follows from our methodology so far is one of a stable structure of ideologies over 20 years and some increase in social liberalism. This increase in social liberalism has occurred both on the intensive margin (ie: the weights on liberal issue-positions in the β vectors) as well as the extensive margin (the growing individual-level type shares for the Liberal Centrist and Left

[86]The type differences are based on the sum of the average type shares for the Liberal Centrist and Left Anarchist types for each country and wave. The figure then plots the difference between a country's average type share in wave 5 and wave 2.

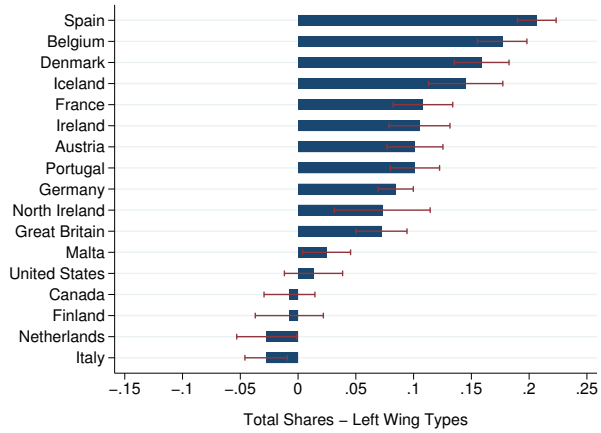
Figure 3.4: Country-Level Type Shares



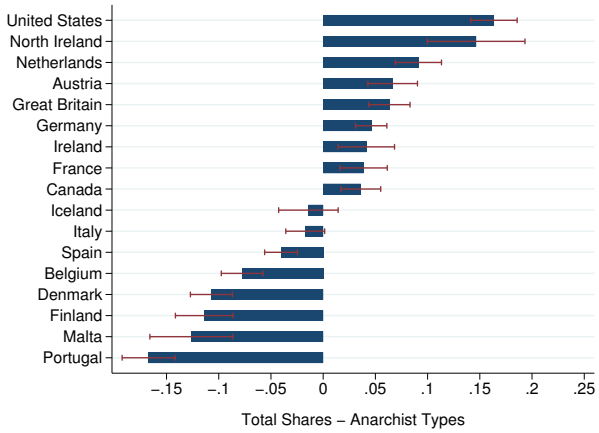
Notes: This figure shows the average country-level θ type shares aggregated over individuals. 95% confidence intervals are reported in red. Country means and confidence intervals are calculated using WVS sample weights. 95% confidence intervals are reported in red.

Figure 3.5: Changes of Types over Time

(a) Change in Left-Wing Types



(b) Change in Anarchist Share

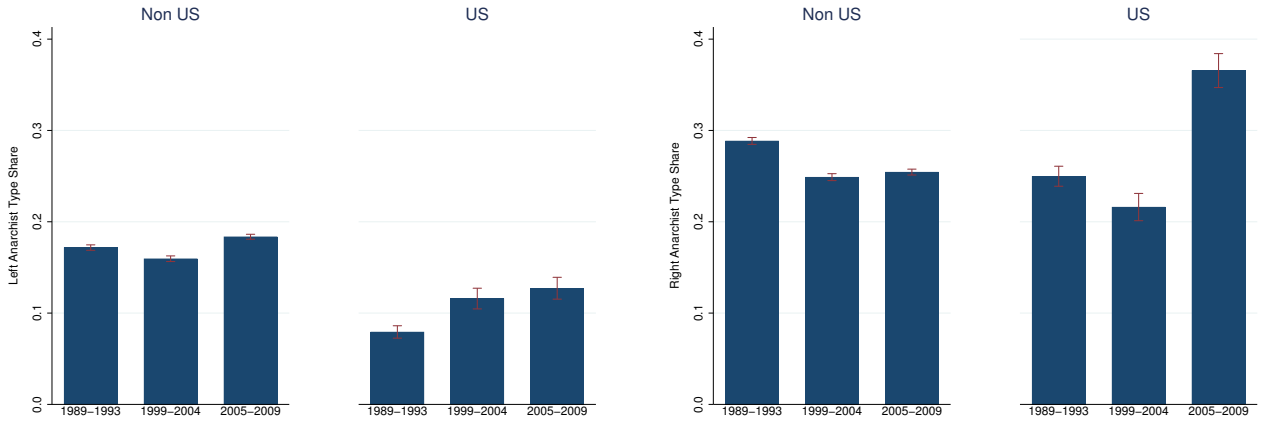


Notes: This figure shows the change in θ type shares by country between Waves 2 (1989-1993) and 5 (2005-2009) in the WVS. In 4(a) we pool the type shares for the Liberal Centrist and Left Anarchist types. In figure 4(b) we show the pooled change in the Left Anarchist and Right Anarchist types. 95% confidence intervals are reported in red.

Figure 3.6: Type Shares - US vs non-US

(a) Left Anarchist Types

(b) Right Anarchist Types



Notes: This figure compares the levels of θ type shares across waves for the Left Anarchist and Right Anarchist types. We pool all 16 non-US countries (effectively all Western European countries apart from Iceland and Canada) and contrast them to the US. The pooling for the non-US sample is based on WVS sample weights. The timing of the waves is Wave 2 (1989-1993), Wave 4 (1999-2004) and Wave 5 (2005-2009). 95% confidence intervals are reported in red.

Anarchist types). The other major development in the data on type shares is the strong tilt towards anti-establishment Anarchist ideologies in some countries - particularly the US.

Ideological Types and the Left-Right Scale.

We next analyze the relationship between our ideological types and the self-positioning of individuals on a left-right scale. For this analysis, we make use of question E033 in the WVS, which asks people to position themselves on a scale of 1 (left) to 10 (right). Recall that this measure of self-positioning is not one of the ingredients in the feature set for our LDA analysis. It is held out from the estimation of the ideological clusters and therefore provides a useful test of validity.

The mean left-right scores according to the dominant type are telling. Individuals with a dominant Left Anarchist type position themselves furthest to the left (mean: 4.33), followed by the Liberal Centrist (5.24), Right Anarchist (5.49) and the Conservative Centrist (mean: 5.74).[87] In line with our previous findings, average political attitudes are moving leftwards with a shift of -0.17 on the left-right scale between waves 2 and 5.

In Figure 3.7, we visualize the relationship between the right-wing (Conservative Centrist and Right Anarchist) type shares and the left-right scale. We find a strong relationship between the type shares and the political orientation of individuals. As expected, the larger the share of the right-wing types in an individual, the further right they place themselves on the political spectrum. The inverse mechanically holds for the left-wing type shares (not shown). This provides further validation for our type labels as they appear to align with the classic left-right ideological spectrum.

Ideological Types and Populism.

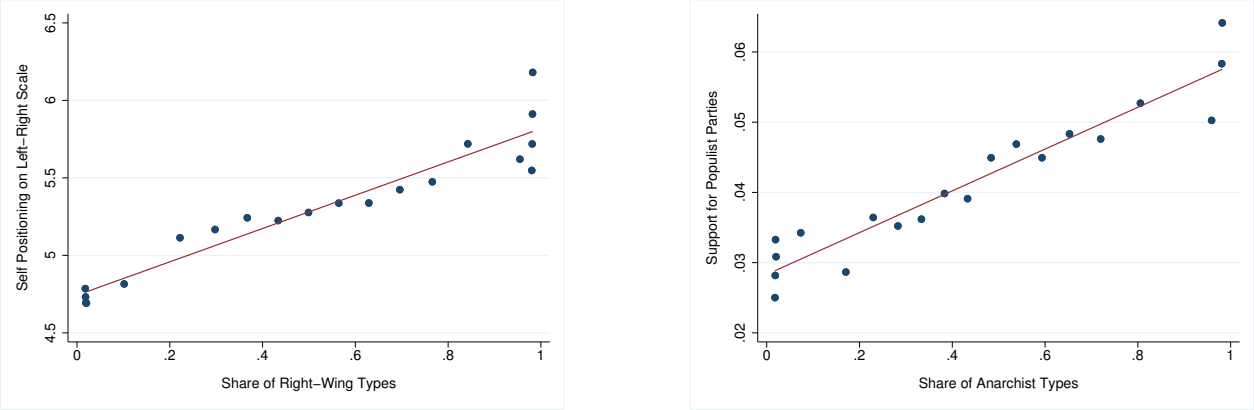
Lastly, we investigate whether our anarchist types are associated with stronger support for populist parties. To do so we consider the question "which political party would you vote for" (question code E179/E179WVS). We recode the responses of individuals as populist parties based on the classification by Rooduijn et al. (2019) (See Section 3.6.9 for additional details). In Figure 3.7 panel (b) we then plot the support for populist parties conditional on the anarchist type shares of the individual. Again a clear positive relationship between voting

[87]Note that, while the mean difference in positioning appears nominally small in these comparisons this is because answers are clustered on middle values: more than 55% of the people position themselves between 4 and 6 on the Left-Right line.

for populist parties and the anarchist ideologies emerges. This suggests that the anarchist types might indeed be a base for populist political mobilisation.

Figure 3.7: Self-positioning on Left-Right Scale and Support for Populist Parties

(a) Self-positioning on Left-Right Scale **(b) Support for Populist Parties**



Notes: The binscatter in panel (a) visualizes the relationship between the individual-level type share of the right-wing types (Conservative Centrist and Right Anarchist) and the self-positioning of individuals on a 1 (left) to 10 (right) scale based on question E033 from the World Value Survey. Panel (b) shows the relationship between the individual-level share of anarchist types and the voting for populist parties coded according to Rooduijn et al. (2019). See Table 3.19 in the online appendix for the full list of parties.

As further evidence, we estimate a simple linear probability model (LPM) of voting for populist parties in Table 3.5, contrasting the explanatory power of our ideological type share variables with that of the left-right self-positioning question. Table 3.5 shows a strong positive relationship between the anarchist type shares and populist voting across all specifications. Interestingly, the left-right self-positioning question only has a significant association when we specify the variable either as a set of dummies for far left or far right positions ((column (4)) or as a step function for each value (Figure 3.8).

In particular, this Figure 3.8 step function shows a U-shape in the probability of populist voting with respect to the left-right scale. That is, people located near the centre of the scale are the least likely to vote populist. One possibility here is that the anarchist type shares are proxying for extreme left or right positioning. However, as noted above, in column (4) of Table 3.5 we control for indicator variables of extreme positioning and this has minimal effects on the previous association. The anarchist type share variables appear to pick up tendencies for populist support from across the left-right spectrum. This association is non-negligible: based on the estimates in Table 3.5, we calculate that an individual with a 50% type share in

Table 3.5: Support for Populist Parties

	(1) Types	(2) L-R Scale	(3) Types L-R Scale	(4) Dummies Far Left/Right
Cons. Centrist	-0.007**		-0.007**	-0.008***
	(0.003)		(0.003)	(0.003)
L. Anarchist	0.037***		0.036***	0.033***
	(0.004)		(0.004)	(0.004)
R. Anarchist	0.033***		0.033***	0.032***
	(0.004)		(0.004)	(0.004)
Left Right Scale		-0.001*	-0.000	
		(0.000)	(0.000)	
I[Far Left]				0.024***
				(0.004)
I[Far Right]				0.028***
				(0.004)
Observations	67,666	67,666	67,666	67,666
R-squared	0.050	0.047	0.050	0.052
Country FE	Yes	Yes	Yes	Yes
Wave FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes

Notes: Each column reports the regression results for individual level regression. The dependent variable is an indicator variable for the support of a populist party. Robust standard errors are used. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1. The data come from the World Value Survey and the European Value Survey. Note that the 67,666 sample presented here is smaller than our main 81,141 sample due to missing values and non-responses for the populist voting and left-right scale questions.

one of the anarchist type has a 34% higher probability of voting for a populist party relative to the sample baseline.[88]

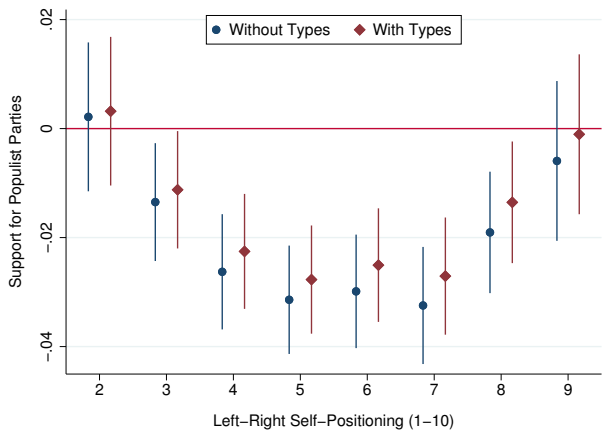
'Citizen Slant' - Within-Person Concentration

Our analysis so far has focused on changes at the level of the ideological types as well as the total shares in the types across the sample. However, for the analysis of the polarisation, the loadings of individuals on the four types is of key importance. In particular, the 'mixed membership' structure of our approach means that two countries with the same overall type distribution can have completely different individual type compositions.

For example, imagine a country which has an overall 50% share in Type 1 and 50% share in a second Type 2. This country can either consist of completely identical individuals with 50% shares in the two types or it could consist of half the population holding a 100% share in Type 1 and another half with a 100% share in Type 2. These two possible type compositions have very different implications for the understanding of societal polarisation. A country

[88] As an additional exercise, we tested whether these findings also held across the entire left-right spectrum by running regressions that split for each value of the left-right scale. For nearly all values of the scale a higher Anarchist type share is associated with more support for populist parties. These results are available by request, though note that the level shift in the U-shape plotted in Figure 3.8 directly corroborates this point.

Figure 3.8: Self-positioning on Left-Right Scale and Support for Populist Parties



Notes: The figure plots the coefficients and 95% confidence intervals for individual level regression, where the dependent variable is an indicator variable for the support of a populist party. The independent variable is a full set of indicators variables for an individuals positioning along the left-right spectrum, the excluded categorie being 1 (far left). The reported coefficients in red additionally controls for the individual level type shares θ_i.

with two separate sets of 'pure' homogeneous types is plausibly more vulnerable to political conflict than a country where there is more ideological heterogeneity at the individual level.

We therefore develop a measure of how strongly an individual is loading on one of the four ideological types by constructing a Gini-style measure of within-person concentration or 'slant'. We define the within-person concentration G_i of individual i as:

$$G_i = \frac{\sum_{t=1}^{4} \sum_{s \neq t} |\theta_i^t - \theta_i^s|}{2(n-1) \sum_{t=1}^{4} \theta_i^t} \tag{3.14}$$

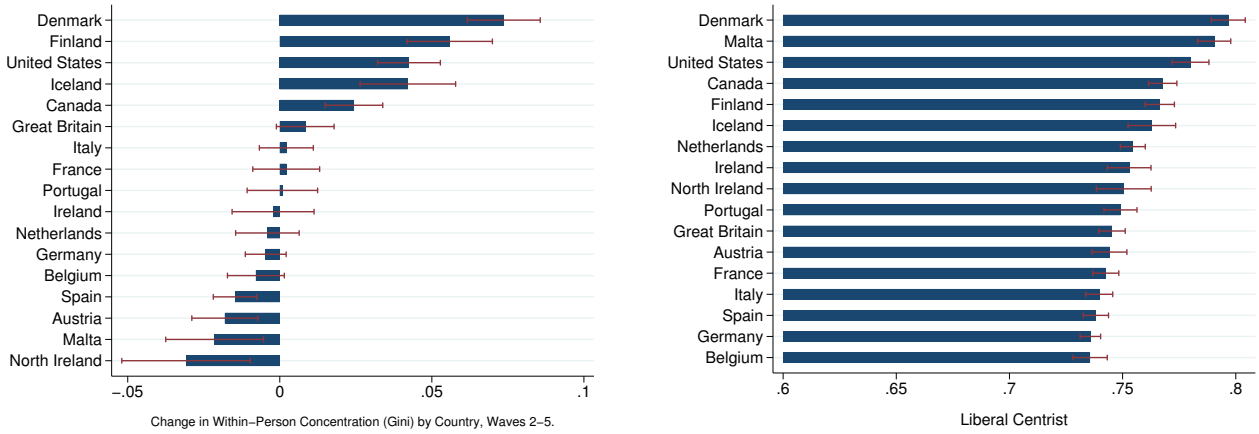
where θ_i^t and θ_i^s are the types shares of individual i. In short, this approach is aggregating the absolute pairwise difference in ideological shares that exist at the individual level. The measure of within-person concentration G_i is monotonically increasing the more an individual loads on one of the ideological types. If a person has a 100% share in one type then G_i will be equal to 1, while $G_i = 0$ implies shares of 25% in all types.

In this way, our measure captures how segregated type shares are on a within-individual basis. Furthermore, it allows us to analyze which groups exhibit a particularly high tendency towards within-person concentration. We plot the (Wave 5) country means of the G_i measure in Figure 3.9 along with the changes between Wave 2 and 5. This shows that G_i is relatively high across the sample with a mean of around 0.75. However, between-country differences are

not dramatic. There is only a 7% gap between the most and least concentrated countries and the ordering does not suggest that any particular ideological types are driving concentration. That is, amongst the most concentrated or 'slanted' countries we see cases of both relatively Conservative and Liberal countries defined in terms of the mean type shares seen earlier. The major, positive country-level shifts in slant over the waves occurred in Denmark, Finland and the US (Figure 3.9a) but the changes were muted for most countries.

Figure 3.9: Citizen Slant by Country

(a) Change in Slant by Country **(b) Ranking of Slant (Wave 5)**



Notes: Panel (a) shows the shows the change in our G_i Gini within-person ideological concentration measure ('citizen slant') from Waves 2 (1989-1993) to 5 (2005-2009) by country. Panel (b) shows the level of the within-person Gini measure by country in Wave 5. 95% confidence intervals are reported in red. Country means and confidence intervals are calculated using WVS sample weights.

To study the importance of individual characteristics on within-person concentration (as well as the development of G_i over time) we estimate the following regression equation:

$$G_{icw} = X_{icw}^{'}\delta + \tau_w + \mu_c + \epsilon_{icw} \tag{3.15}$$

where X_{icw} is a vector of covariates[89], τ_w are wave dummies, μ_c are the country dummies and G_{icw} is the of Gini coefficient of individual i in country c and wave w. The results are reported in Table 3.6 with controls for the type shares and with the Liberal Centrist set as the baseline type. The purpose of the type share controls is to allow us to study whether G_i concentration is increasing with shares of certain types. The results indicate that the Left Anarchist is the least concentrated type followed by the Right Anarchist. In turn, this means

[89]We use the same individual covariates as in Table 3.4.

that the individuals with larger shares in either of the two anarchist ideologies are more likely to mix different ideological types than the centrist types.

After controlling for the available individual characteristics we find a 1.6% increase in G in wave 4 and an 0.6% increase in wave 5[90]. The results for the analysis of the US are similar overall except that the increases of G concentration in Waves 4 and 5 are far larger, standing at 2.8% and 5.0% (Column 3).

To further probe the increases in G over time we estimate eq. (3.15) separately for individuals conditional on their main type and also broken down according to the US and non-US samples. The results in Table 3.7 show that the increases in G within the US are predominantly driven by the two Anarchist types, both of which exhibit increases in concentration around 13% relative to the overall sample mean (0.753).

Table 3.6: Correlates of 'Citizen Slant' (Gini Concentration)

	(1) Gini	(2) Gini	(3) Gini US	(4) Gini US
Cons. Centrist	0.000		-0.013**	
	(0.002)		(0.006)	
L. Anarchist	-0.035***		-0.060***	
	(0.002)		(0.010)	
R. Anarchist	-0.024***		-0.045***	
	(0.002)		(0.007)	
Wave 4	0.016***	0.016***	0.028***	0.028***
	(0.002)	(0.002)	(0.006)	(0.006)
Wave 5	0.006***	0.006***	0.050***	0.042***
	(0.001)	(0.001)	(0.006)	(0.006)
Observations	81,141	81,141	4,197	4,197
R-squared	0.019	0.010	0.037	0.017
Controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Mean Dep. Variable	0.753	0.753	0.759	0.759

Notes: Each column reports the regression results for individual level regression. The dependent variable is the Gini Coefficient of the individual type shares as a measure of polarisation. Column (1) and (2) use all data and column (3) and (4) restrict the sample to the USA. Robust standard errors are used. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1. The data come from the World Value Survey and the European Value Survey.

To clarify, note that the message from the earlier table was that the Anarchist types are less concentrated in the cross-section (hence the positive coefficients on these type variables in the associated regressions). In contrast, the regressions in Table 3.7 track how concentration developed over time on a per-type basis. The simple story then is that, when they do manifest,

[90]We suppress the reporting of the individual attribute coefficients in eq. (3.15) to avoid clutter. The basic result for these covariates is that only gender (female) and unemployment contribute significantly to within-person concentration but with small magnitudes. They enter with positive and negative signs respectively.

Anarchist views are becoming more concentrated or 'purer' at the individual level rather than being spread out amongst more people.

In effect, this evidence implies that the Anarchist types have become an even more dominant ideology for people who had already shown a lack of trust in social and political institutions. While in earlier waves this section of the population might still have shown large type shares in Centrist ideologies this potentially moderating centrist influence became less evident in more recent years. The findings for the US also contrast fairly strongly with the results for the non-US sample where the increase in concentration for the Anarchist types is more muted and, in any case, is matched by increases for the Liberal Centrist type as well (see Table 3.7, panel (B), first column).

3.4.5 Societal Polarisation

While the above measure of within-person concentration describes the strength of the individual loadings on the four ideological types, it does not fully summarise the extent of the divisions between citizens within a society. A society in which there are sub-groups of individuals that heavily load on the same ideological type may not necessarily be dramatically polarised. The extent of polarisation would hinge on how big these 'purist' sub-groups are relative to the full set of ideological sub-groups across the population. As an example, the country-level type share plots in Figure 3.4 indicate that some countries are characterised by widely represented types with aggregate type shares around the 50% mark, such as Liberal Centrists in Denmark and Conservative Centrists in Malta. At face value, these countries could be plausibly classified as 'unipolar' and less vulnerable to group conflict no matter how concentrated the different types are in terms of citizen slant.

We, therefore, study polarisation by adapting the measures proposed by Esteban and Ray (1994) and Duclos et al. (2004) to our setting with 4 ideological types. These measures have the feature of accommodating two forces that define polarisation as a general concept. Firstly, there is *identification* which occurs amongst individuals with a common characteristic and is an increasing function of the total number of common individuals (that is, group size). Secondly, there is *alienation* which accounts for the social detachment that individuals feel towards others who do not share some common characteristic. Again, the strength of the alienation effect will depend on (relative) group size as well as the 'distance' between groups in the key characteristic of concern.

Table 3.7: 'Citizen Slant' - US vs non-US Comparison

Panel A: United States				
	(1) Lib. Centrist	(2) Cons. Centrist	(3) Left Anarchist	(4) Right Anarchist
Wave 4	0.036***	0.008	0.037	0.046***
	(0.009)	(0.010)	(0.024)	(0.012)
Wave 5	0.006	0.040***	0.098***	0.095***
	(0.012)	(0.012)	(0.023)	(0.011)
Observations	1,412	1,408	267	1,110
R-squared	0.016	0.018	0.090	0.090
Controls	Yes	Yes	Yes	Yes
Panel B: Non United States				
	(1) Lib. Centrist	(2) Cons. Centrist	(3) Left Anarchist	(4) Right Anarchist
Wave 4	0.045***	-0.007**	0.014***	0.006**
	(0.003)	(0.003)	(0.004)	(0.003)
Wave 5	0.029***	-0.013***	0.004	0.001
	(0.003)	(0.003)	(0.004)	(0.003)
Observations	24,339	21,326	10,766	20,513
R-squared	0.055	0.041	0.032	0.022
Controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Panel C: All Countries				
	(1) Lib. Centrist	(2) Cons. Centrist	(3) Left Anarchist	(4) Right Anarchist
Wave 4	0.044***	-0.005*	0.015***	0.009***
	(0.003)	(0.003)	(0.004)	(0.003)
Wave 5	0.028***	-0.010***	0.006*	0.007***
	(0.003)	(0.003)	(0.004)	(0.003)
Observations	25,751	22,734	11,033	21,623
R-squared	0.053	0.038	0.032	0.021
Controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes

Notes: Each column reports the regression results for individual level regression. The dependent variable is the Gini Coefficient of the individual type shares as a measure of polarisation. Column (1) use all US data and column (2), (3) and (4) restrict the sample to the individuals based on their dominant type. Robust standard errors are used. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1. The data come from the World Value Survey and the European Value Survey.

Using the example of income inequality, Esteban and Ray (1994) prove that any measure of polarisation P that accurately accounts for own-group identification as well as alienation in relation to an out-group and fulfills 3 'reasonable' axioms has to be of the form[91]:

$$P(\pi, y) = \kappa \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i^{1+\nu} \pi_j |y_i - y_j| \tag{3.16}$$

where π are the number of people in the groups, y is amount of income of each group, K is a normalizing constant and ν is the polarisation sensitivity, which parameterises how the polarisation measure shifts with group sizes.

This general polarisation measure P was constructed for a one dimensional variable y, for example, income. Polarisation in our case has to be measured over all 4 ideological types. To do this, we divide people into meaningful ideological groups based on their dominant type share. That is, we allocate individuals to one of our 4 groups based on their highest type share at the individual i-level. We then add up the θ_i type shares amongst the defined group members to get the mean type share, defined as $\tilde{\theta}_t$. This differs from the mean type shares $\bar{\theta}_t$ we have presented earlier on the basis that we are only taking the mean over individuals with the *same dominant type* rather over the whole population.

Given this modification, our polarisation measure is defined as:

$$P(\pi, \theta) = \kappa \sum_{t=1}^{4} \sum_{j=1}^{4} \pi_t^{1+\nu} \pi_j (|\tilde{\theta}_{t1} - \tilde{\theta}_{j_1}| \rho_{t1} + |\tilde{\theta}_{t2} - \tilde{\theta}_{j_2}| \rho_{t2} + |\tilde{\theta}_{t3} - \tilde{\theta}_{j_3}| \rho_{t3} + |\tilde{\theta}_{t4} - \tilde{\theta}_{j_4}| \rho_{t4})$$

where π_t and π_j are the number of people who have the dominant type share t and j. The means of the type shares in the each of the 4 dominant type groups are $\tilde{\theta}_t$ for own type and $\tilde{\theta}_j$ for a generic other type. The second subscript on $\tilde{\theta}_t$ and $\tilde{\theta}_j$ represents the dominant type group we are conditioning on when calculating the absolute distance between groups. Finally, $\rho_{tj} = \frac{3 - corr(\beta_t, \beta_j)}{2}$ uses information from the β type vectors. As such, ρ_{tj} is a measure of the similarity of types based on the correlation of types rescaled to be contained in the $[1, 2]$ interval. Individuals of dominant type t weight differences in type t by 1 while all other type differences have weight strictly larger than one.

As an example, consider setting type t as the Liberal Centrist and j is the Conservative Centrist. We index the Liberal Centrist as the type 1 in the second conditioning subscript. The calculation $|\tilde{\theta}_{t1} - \tilde{\theta}_{j_1}|$ then represents the (absolute) difference between the mean Liberal

[91] The Axioms put forward in Esteban and Ray (1994) are explained in more detail in Appendix 3.6.10.

Centrist type share for dominant Liberal Centrist individuals and the mean Liberal Centrist type share for dominant Conservative Centrist individuals. This can be interpreted as a measure of how close different ideological groups are despite their contrasting dominant type shares. That is, a Liberal Centrist and a Conservative Centrist are more likely to 'get along' if they have high minority type shares in each other's ideology.

The other components of $P(\pi, \theta)$ are the polarisation sensitivity parameter ν, which we fix at $\nu = 0.5$, and the constant $\kappa = (\sum_{t=1}^{4} \pi_t)^{-(2+\nu)}$ that serves to normalize the polarisation measure by population size. We provide more detail and show how P varies with different values of ν in Appendix 3.6.10.

Intuitively, the polarisation measure will be largest for the case where there are two major type share groups of identical size who exhibit completely different type shares. An example of this would be a bipolar Liberal Centrist and Right Anarchist society where each type group had very small minority shares in the other type. This provides a natural link back to our earlier measure of citizen slant. Since an increase in citizen slant implies an increase in the means for $\tilde{\theta}_t$ and $\tilde{\theta}_j$, absolute differences in type shares between the groups increase and polarisation rises due to stronger alienation effects.

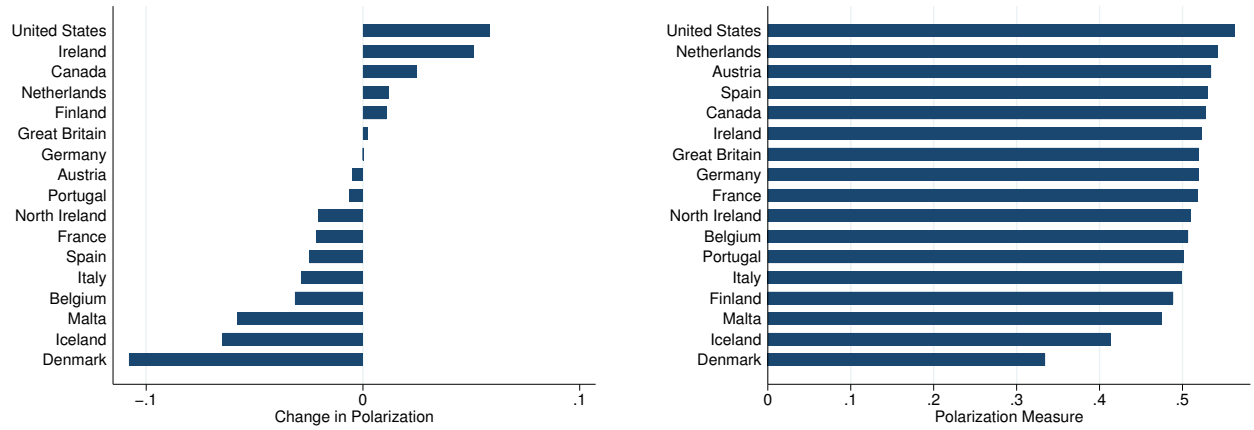
It is also useful to note how polarisation also depends on the relative sizes of the groups within a population, as measured by π_t and π_j. For example, given the same between-group differences in types, a country in which 2 groups each make up 50% of the population will be more polarised than a country with 4 groups each making up 25% of the population.

We calculate the polarisation measure separately for each country and wave in our sample. The ranking of the countries based on their polarisation in each wave is shown in Figure 3.10. The ranking of countries according to Wave 5 polarisation levels is distinct from the earlier ranking for citizen slant. Denmark, which has the lowest level of polarisation, provides an instructive example of how the P polarisation measure combines information. The inputs into the result for Denmark are its high Liberal Centrist type share (above 0.5 - see Figure 3.4) and high level of within-person concentration or slant (which intensifies over time - see Figure 3.9). Hence the low Danish P measure reflects a case of ideological consensus where there is a major plural type (Liberal Centrist) that is strongly held by individuals (as manifested in high slant).

The US, which sits at the top of the polarisation ranking in Wave 5, provides a sharp comparison that again illustrates the mechanics of the P measure. It has a relatively even spread of type shares, with shares of around 30% for the Liberal Centrist, Conservative

Figure 3.10: Polarisation by Country

(a) Change in polarisation by Country **(b) Ranking of polarisation (Wave 5)**



Notes: Panel (a) shows the change in country-level polarisation measures from Waves 2 (1989-1993) to 5 (2005-2009) calculated following Esteban and Ray (1994). Panel (b) shows the level of the country-level polarisation measure in Wave 5.

Centrist and Right Anarchist types. Hence, the group size effect picked up by the π_t and π_j terms is stronger in the US compared to unipolar cases such as Denmark. Overall, the increase in polarisation in the US is mainly driven by the rise in slant over time (Figure 3.9), which contributes by intensifying the alienation effect. The changes in dominant type composition in the US only have a minor influence on the polarisation measure.

However, it should be noted that, across countries, the changes in polarisation over time are not dramatic, with most of the shifts occurring in the 2-5% range relative to baseline values in Wave 2. A key point to note is that the defining feature of some of the polarisation episodes seen in the data is the *qualitative* content of developments. The US is the banner example here since the increase in polarisation was driven by an increase in the presence of Anarchist types. Hence the US experience with polarisation has the extra dimension of also reflecting the trend of a 'disappearing centre', which is likely to have additional consequences for social cohesion over and above the increase in P that we measure statistically.

In unreported results, we also calculated our polarisation measure using the 3, 5, 6 and 7-type models. The US is consistently either the top country or within the top 3 of most polarised countries across these models.

3.5 Conclusion

In this paper, we have proposed a new way to identify the latent ideologies of individuals from survey data. Our approach does not presuppose any ideological structure for individuals. Nonetheless, we are able to identify interpretable, consistent and stable ideological types in the data. The findings generally align with the left-right framing frequently used in the social sciences but we also identify anti-establishment 'anarchist' ideologies that are characterized by their distrust in important societal institutions.

The approach taken in this paper can also be extended in a number of directions. Firstly, the basic approach outlined here can be applied to other survey datasets, both for the countries studied here and for those outside North America and Western Europe. Indeed, our measure of topic cohesion might be used for any topic modelling application with non-text data. Secondly, the approach is general enough to be used to study questions beyond political ideologies, such as clusters of cognitive and non-cognitive skills, behavioral patterns or personality types.

This latter point about extensions that cover subjects apart from political views is potentially very rich. Contributions such as Ortoleva and Snowberg (2015); Chapman et al. (2018) and Enke (2020) have identified behavioural foundations of political views that systematically map into voting and other outcomes. However, we see it as plausible that our low trust 'anarchist' types may have some underpinnings in a further layer of personality or behavioral characteristics. Given sufficiently rich data, these layers could be modelled and validated using the hierarchical, out-of-sample approach we have outlined in this paper. Simon's (1969) concept of a *hierarchic system* is a potential model for such future work and we think that our paper shows the potential of unsupervised learning methods to model such latent, unobserved characteristics at a new level of depth.

3.6 Appendix: How Polarized are Citizens

3.6.1 Appendix: Additional Details on the Selection of Question from the WVS

This section describes in more detail, the selection process that lead to the 29 questions that are used in the paper. There are 6 waves of the World Value Survey (WVS) and 4 waves of the European Value Survey (EVS). The 4 Waves of the EVS correspond to the 1st, 2nd, 4th and 5th wave of the WVS. When constructing the Integrated Value Survey by combining the WVS and EVS we excluded the 1st wave since, it contained a smaller set of countries and questions. The Integrated Value Survey (WVS) in total contains 971 different items grouped in 13 different categories (number of questions in brackets): Environment (25), Family (64), National Identity (105), Perceptions of life (210), Politics and Society (267), Religion and Morale (122), Science (2), Security (22), Socio-demographics (38), Special Indexes (3), Structure of the file (25), Sylatech module (42) and Work (46).[92]

We limited the set of questions to those question which were consistently asked in the 2nd, 4th and 5th wave of WVS. This already reduced the set of possible questions down to 92. From these 92 questions we chose our 29 based on which questions are most salient for the evaluation of a person's ideological type. The excluded question are listed in Table 3.8. For example, we exclude questions about family structure (eg: single parenting, beliefs in marriage), questions about non-political moral values (eg: 'important child qualities'), life satisfaction, and generic trust in others. This is because our aim is to model the latent structure of political opinions that are most analogous to the concept of ideology. In the conclusion of the paper we describe possible extensions of our general approach that would accommodate an interaction between (say) behavioral characteristics and political beliefs.

In Section 3.6.5 we show that the selection of these 29 questions is not crucial for our findings and that the ideological types are very similar if we use all 92 questions. We further show that also removing any of the 29 questions from our data has no bearing on our results.

A further point is that LDA does not allow for missing responses in the data. If we simply excluded all observations with any missing responses and restricted ourselves to observations with complete sets of answers, we would need to drop sizable fractions of the WVS data. We instead impute a small set of missing responses with the sample mean of the non-missing

[92]The categories socio-demographics, special indexes, structure of the study, Sylatech module and work do not contain any question concerning the values of people.

Table 3.8: List of Excluded Questions

Code	Questions	Code	Questions
A001	Important in life: Family	D057	Being a housewife just as fulfilling
A002	Important in life: Friends	E001	Aims of country: first choice
A003	Important in life: Leisure time	E002	Aims of country: second choice
A004	Important in life: Politics	E003	Aims of respondent: first choice
A005	Important in life: Work	E004	Aims of respondent: second choice
A006	Important in life: Religion	E005	Most important: first choice
A008	Feeling of happiness	E006	Most important: second choice
A009	State of health (subjective)	E012	Willingness to fight for country
A029	Important child qualities: independence	E015	Future changes: Less importance placed on work
A030	Important child qualities: hard work	E016	Future changes: More emphasis on technology
A032	Important child qualities: feeling of responsibility	E018	Future changes: Greater respect for authority
A034	Important child qualities: imagination	E019	Future changes: More emphasis on family life
A035	Important child qualities: tolerance and respect for other people	E022	Opinion about scientific advances
A038	Important child qualities: thrift saving money and things	E023	Interest in politics
A039	Important child qualities: determination perseverance	E025	Political action: signing a petition
A040	Important child qualities: religious faith	E026	Political action: joining in boycotts
A041	Important child qualities: unselfishness	E027	Political action: attending lawful/peaceful demonstrations
A042	Important child qualities: obedience	E033	Self positioning in political scale
A124_03	Neighbours: Heavy drinkers	E035	Income equality
A124_05	Neighbours: Muslims	E069_11	Confidence: The Government
A165	Most people can be trusted	E069_12	Confidence: The Political Parties
A170	Satisfaction with your life	E069_18	Confidence: The European Union
A173	How much freedom of choice and control	F001	Thinking about meaning and purpose of life
B001	Would give part of my income for the environment	F028	How often do you attend religious services
B002	Increase in taxes if used to prevent environmental pollution	F034	Religious person
B003	Government should reduce environmental pollution	F035	Churches give answers: moral problems
C001	Jobs scarce: Men should have more right to a job than women	F036	Churches give answers: the problems of family life
C006	Satisfaction with financial situation of household	F037	Churches give answers: people´s spiritual needs
C059	Fairness: One secretary is paid more	F038	Churches give answers: the social problems
D018	Child needs a home with father and mother	F063	How important is God in your life
D022	Marriage is an out-dated institution	F065	Moments of prayer, meditation...
D023	Woman as a single parent		

Notes: This table contains the questions that were excluded from baseline LDA model.

data in the same wave. This treatment of missing data allows LDA to use the information from this observation across other questions that have non-missing values. Moreover, the imputation has only a minimal effect on the LDA classification, since the sample mean does not influence the classification of each individual. Imputation with the mean is also preferable to an alternative approach where we would simply replace all missing responses with 0s, because the 0s would bias the classification.

In Table 3.14 we report the resulting type hierarchy for the approximate 50% of observations in the sample that do not require any imputation. As it turns out the resulting type hierarchy is nearly identical to the one achieved with imputation. Also the resulting individual-level type shares are very similar. This suggest that the imputation of missings is, as expected, not having a major influence on our results.

3.6.2 Appendix: Additional Details on the LDA Model Inference

One difference between our application and the standard use of LDA is that in our case features can only appear once for each observation, i.e. people can only answer each question once, while words can appear more than once in a document. As a result, the assumption of the independence of features is violated in the case of the issue positions from the WVS. This section describes while the LDA model inference nonetheless remains valid and compares LDA to the model suggested by Gross and Manrique-Vallier (2012).

To understand why the approximation of the likelihood works even when features can only appear once, it is helpful to analyze the updating steps of the approximation algorithms. Of the existing LDA approximation algorithms the collapsed Gibbs Sampling algorithm developed by Griffiths and Steyvers (2004) and for example used in Schwarz (2018) provides the clearest inside into the workings of LDA. The collapsed Gibbs sampler works by consecutively sampling new topic (type) assignment for each feature – words in text or individual's question responses –, based on the current topic assignment of all other features.[93]

In our application the Gibbs Sampler would calculate the probabilities for $z_{i,n}$ – the type assignment of response n and individual i – conditional on $z_{-(i,n)}$ the current type of all other features and the given response r based on the following equation:

$$P(z_i = t|z_{-r}, \beta) \propto \frac{\eta_t^{(r)} + \gamma}{\eta_t^{(.)} + Q\gamma} \cdot \frac{n_i^{(t)} + \alpha}{n_i^{(.)} + T\alpha}$$

[93]In the beginning, the Gibbs Sampler is initialized by randomly assigning features to topics (types)

, where $n_t^{(r)}$ is the number of times response r is currently assigned to type t and $\eta_t^{(.)}$ is the number of times any response is assigned to type t. Similarly, $\eta_t^{(i)}$ is the number of responses of individual i assigned to type t and $\eta_i^{(.)}$ is the total number of responses given by individual i. α, γ are the Dirichlet priors, Q is the number of Questions (58 in our case) and T is the number of types. The first fraction hence captures the probability to observe response r conditional on type t, while the second fraction captures the probability of type assignment t for individual i.

After calculating $P(z_i = t|z_{-r}, w)$ for all $t \in T$, the Gibbs sampler randomly draws a new type assignment based on the calculated probabilities. In other words, the Gibbs sampler exploits how likely $P(z_i = 1|z_{-r}, w)$ is relative to $P(z_i = 2|z_{-r}, w), \cdots, P(z_i = T|z_{-r}, w)$. Hence, assignment of $z_{i,q}$ to types captures the relative frequency of response r conditional on types. Since all individuals can give response r at most once the type assignments are valid. The type assignment only would be biased if individuals would differ in how often they could give response r.

Gross and Manrique-Vallier (2012) developed an alternative model to model survey responses, which leads to an alternative updating equation:

$$P(z_i = t|z_{-r}, \beta) \propto \frac{\eta_t^{(r)} + \gamma}{\eta_t^{(\neg r)} + R\gamma} \cdot \frac{\eta_i^{(t)} + \alpha}{\eta_i^{(.)} + T\alpha}$$

, where $\eta_t^{(\neg r)}$ is the number of times another response is given to question q and R is the number of possible responses to question q. Therefore, the difference between the two updating equations is that LDA uses the probability of response r relative to all other responses given by type t, while the Gross and Manrique-Vallier (2012) model uses the probability relative to other responses given to the same question. Hence, in LDA the probabilities of responses sum to one across all questions ($\sum_{q=1}^{Q} \beta_q = 1$). Whereas in the (Gross and Manrique-Vallier, 2012) model the probabilities of responses sum to one for each individual question ($\sum_{r=1}^{R} \beta_{q,r} = 1$).

Put more simply, in the Gross and Manrique-Vallier (2012) model every question is treated separately. In the LDA model putting weight on one issue reduces the weight on other issues. In this way, the LDA model creates a natural hierarchy of issue positions and their importance for the ideological types. While the model of Gross and Manrique-Vallier (2012) arguably better describes the actual data generation process our LDA approach better accounts for the fact that ideological types might focus on a group of issue positions, e.g. social issues, while the responses to other questions, e.g. economic issues, might be less

important. Our recoding of questions into 2 features also naturally incorporates this feature of LDA, as an individual who for example states trust in the government cannot also state distrust in the government.

3.6.3 Appendix: Interpretation of the β Vectors

LDA allows for repeated draws of a feature, while in our application people can only answer a question once. As already discussed in the main part of the paper and Section 3.6.2 this does not influence on the validity of LDA, since LDA exploits how often features appear relative to each other.

However, this difference influences the interpretation of the β vectors. The β vectors capture the probability that a response is drawn in each of the 29 draws (questions) asked to an individual, e.g. how likely it is that an individual will answer that he is opposed to abortion in each of the 29 draws. Therefore, the β vectors do not take into account that once a person has answered a question the same person cannot answer the same question again.

As a result, the β still capture which groups are more likely to exhibit an ideological position, but the values do not have a natural interpretation within our setting. If necessary one can scale up the β probabilities to give them a more natural interpretation within our setting. To do this one has to calculate the probability that a feature shows up in any of the 29 draws of the LDA taking into account that a question can only be answered once. Given this intuition $P_{f,t}$, the overall probability that a feature f appears if the chosen type is t, can be expressed as $P_{f,t} = \sum_{d=1}^{29}(1-\beta_{f,t})^{d-1}\beta_{f,t}$, where d is the number of the draw (question) and $\beta_{f,t}$ is the value of the β vector for feature f and type t. In this expression $(1-\beta_{f,t})^{d-1}$ is the probability that the response has not been given in any previous draw and $\beta_{f,t}$ is the probability that the response will be given in the current draw.

As an example to illustrate this calculation consider the question of "Confidence in the Police". In the 5th wave, the liberal centrist has a value of $\beta_{14,1} = 0.0408$ and the value for the left anarchist is $\beta_{14,3} = 0.0089$. This difference in the β values translate into the following overall differences in probability. While a liberal centrist will express confidence in the police with a probability of 70.1%, the probability that a left anarchist will express similar views is only 22.8%.

This scaling-up does not take into account some features are mutually exclusive. Hence, the scaled-up probability of the features "Confidence in the Police" and "No Confidence in

the Police" will not necessarily add up to 1.

3.6.4 Appendix: Additional Details on Topic Cohesion

Automatic Evaluation of Topic Model Cohesion

The main theme of the literature on the cohesion of topic models is that humans judge topics to be more consistent based on word co-occurrences (Chang et al., 2009; Newman et al., 2010; Lau et al., 2014; Lau and Baldwin, 2016). Consider, for example, a topic containing words like 'labour', 'wage' and 'firm', which often appear together in a text, will be judged as highly coherent by humans. An alternative topic that contains words like 'inflation', 'agriculture' and 'hospital' appears incoherent since these words are not used together as frequently.

Given this approach, it is possible to automatically calculate measures of topic cohesion that are highly correlated with human judgment. These measures are usually based on the most frequently occurring words in each topic. One standard approach is to calculate how often words appear together using the Wikipedia corpus (Newman et al., 2010). The title and sub-sections of the Wikipedia article are used as 'tags' for discrete, human-curated topics. The more frequently that words within an LDA-derived topic appear together in a Wikipedia article (or within a sub-section of an article) then the more coherent the automatically defined topic is judged to be.

In our specific case of using survey response data, there is no equivalent, human-curated outside corpus available to guide analysis. We, therefore, take the approach of using hold-out samples from within our data to calculate the cohesion scores. Our method thereby exploits the same intuition normally used in the literature on topic model cohesion. The key here is the β issue-position weights can be used as predictions of feature co-occurrence in the hold-out data. A political ideology is judged to be more coherent, if people frequently hold issue-position's together. We use Normalized Pointwise Mutual Information (NPMI) as our score of topic cohesion since NPMI has been shown to outperform other information metrics such as PMI or Pairwise Log Conditional Probability (LCP) and shows similar performance to pairwise distributional similarity (Aletras and Stevenson, 2013; Lau et al., 2014).

Making Sense of the NPMI Values

The calculation of the NPMI is based on the independent and joint probabilities of given features k and l. The probability $p(k)$, for example, could capture the share of the population

that believes abortion is not justifiable, while $p(l)$ captures the probability that a person has confidence in the church. The joint probability $p(k, l)$ then captures how many people believe that abortion is unjustifiable and have confidence in the church at the same time. The larger the joint $p(k, l)$ is in relation to $p(k)$ and $p(l)$, the higher is NPMI score of the two features.

Re-capping the basic equation from the main paper NPMI is defined as:

$$NPMI_{k,l} = \frac{PMI_{k,l}}{-\ln(p(k,l))} = \frac{\ln\left(\frac{p(k,l)}{p(k)\cdot p(l)}\right)}{-\ln(p(k,l))} \tag{3.17}$$

As an illustration, Table 3.9 shows two examples of NPMI scores for different values of $p(k)$ and $p(l)$, as well as different joint probabilities $p(k, l)$. In the first example, both features appear with a probability of 0.2. In the situation where all people who are against abortion also have confidence in the church, the joint probability of the features is 0.2 and the NPMI value will be 1. If the two features were independent of each other one would expect them to appear together in the data with a frequency of $(0.2 \cdot 0.2) = 0.04$. In this situation, the calculated NPMI will be 0. If the joint probability is larger than the probability in the case of independence then NPMI will be positive, with the converse applying. The final two rows of Example 1 in Table 3.9 illustrate this relationship.

Table 3.9: Example Calculation NPMI

Example 1					
Case	$p(k)$	$p(l)$	$p(k,l)$	PMI	$NPMI$
Perfect Co-Occurrence	0.2	0.2	0.2	1.609	1
Independence	0.2	0.2	0.04	0	0
$p(k,l) > Independence$	0.2	0.2	0.06	0.405	0.244
$p(k,l) < Independence$	0.2	0.2	0.02	-0.693	-0.177
Example 2					
Perfect Co-Occurrence	0.6	0.6	0.6	0.511	1
Independence	0.6	0.6	0.36	0	0
$p(k,l) > Independence$	0.6	0.6	0.54	0.405	0.658
$p(k,l) < Independence$	0.6	0.6	0.18	-0.693	-0.404

A technical point to note here is that the exact value of the NPMI depends on the individual as well as the joint probabilities. This is illustrated via the second example reported in Table 3.9. Note that in both Example 1 and Example 2 the third row cases are characterized by a joint probability that is 50% larger than in the case of independence. The PMI is identical across the two different 'third row' cases but the NPMI is different. Two pairs of feature

will only have the same NPMI if $log_{p(k,l)}(p(k), p(l)) = log_{p(k,l)}(p(k), p(l))$. In other words, the NPMI is identical, if you have to raise the joint probability to the same power to get the product of the individual probabilities.

3.6.5 Appendix: Sensitivity to Removal and Addition of Features

In this section, we analyze how sensitive our baseline 4-type model is to the removal and addition of features. The exercises we run here can be interpreted as a leverage or influence analysis on the statistical definition of our ideological clusters. We are unaware of formal model robustness statistics of this nature in the literature on LDA. Hence, while we think that the exercises below are promising in terms of the robustness of the basic clusters that they reveal, they should be considered indicative.

'Leave One Out' Clusters.

As a first exercise, we re-estimate the 4-type model removing 1 of the 29 questions (2 of the 58 features) at a time. Afterwards, we compare the original model to the new 'leave one out' model based on the similarity of the β vectors, as measured by their correlation. Table 3.10 reports the results of this exercise.

Overall, independent of the particular removed question, we find high correlations between the different β vectors. This is strongest for the Liberal Centrist type which has an average correlation of 0.979 between the original and leave one out models across all dropped questions. This indicates that the types generated by LDA are very closely comparable across the different models. The highest degrees of sensitivity relate to the confidence in institutions questions (where the β correlations are between 0.70-0.80 for three of the types). Another point of sensitivity is questions relating to foreigners/immigration in the case of the Right Anarchist. Given the centrality of the confidence and immigration questions to the character of different types, these sensitivities are within expectations. This leads to the next issue of how the types might change when we add more information into the feature set.

Widening the Feature Set.

In the next exercise, we investigate how the structure of our clusters changes when we include additional features in the topic model. As described in Section 3.6.1, there are a total of

Table 3.10: Sensitivity Removal of Features - 'Leave One Out' Exercise.

Question Code	Removed Question	Type 1 Lib. Centrist	Type 2 Cons. Centrist	Type 3 Left Anarchist	Type 4 Right Anarchist
A124_02	Against Neighbours: People different race	0.990	0.973	0.995	0.963
A124_06	Against Neighbours: Immigrants/foreign workers	0.991	0.976	0.996	0.955
A124_07	Against Neighbours: People AIDS	0.989	0.958	0.997	0.903
A124_08	Against Neighbours: Drug addicts	0.999	0.997	0.997	0.947
A124_09	Against Neighbours: Homosexuals	0.988	0.824	0.959	0.448
	Average Neighbours:	0.992	0.945	0.987	0.843
C002	If Jobs scarce: priority to (nation) people	0.999	0.995	0.995	0.928
E036	Private better than state ownership	0.999	0.994	0.995	0.923
E037	More responsibility for government	0.999	0.998	0.997	0.957
E039	Competition is good	0.998	0.998	0.997	0.958
	Average Economics:	0.999	0.996	0.996	0.941
E069_01	Confidence: Churches	0.993	0.836	0.954	0.577
E069_02	Confidence: Armed Forces	0.983	0.817	0.959	0.627
E069_04	Confidence: Press	0.976	0.833	0.951	0.727
E069_05	Confidence: Labour Unions	0.982	0.799	0.959	0.645
E069_06	Confidence: Police	0.976	0.951	0.980	0.284
E069_07	Confidence: Parliament	0.990	0.975	0.982	0.276
E069_08	Confidence: The Civil Services	0.990	0.975	0.983	0.289
E069_13	Confidence: Major Companies	0.993	0.966	0.979	0.135
E069_17	Confidence: Justice System/Courts	0.933	0.851	0.944	-0.340
	Average Confidence	0.980	0.889	0.966	0.358
F114	Justifiable: claiming government benets	0.999	0.998	0.997	0.958
F115	Justifiable: avoiding a fare on public transport	0.645	0.750	0.940	0.702
F116	Justifiable: cheating on taxes	0.999	0.998	0.997	0.955
F117	Justifiable: accepting a bribe	0.999	0.998	0.998	0.959
	Average Fairness Values	0.911	0.936	0.983	0.894
F118	Justifiable: homosexuality	0.988	0.993	0.992	0.970
F119	Justifiable: prostitution	0.999	0.999	0.998	0.974
F120	Justifiable: abortion	0.999	0.999	0.999	0.974
F121	Justifiable: divorce	0.994	0.995	0.997	0.969
F122	Justifiable: euthanasia	1.000	0.998	0.999	0.981
F123	Justifiable: suicide	1.000	0.999	1.000	0.986
	Average Social Values	0.996	0.997	0.997	0.975
G006	Proud of nationality	1.000	0.998	0.999	0.985
	Average All:	0.979	0.946	0.984	0.745

Notes: This table reports the correlation of the β vectors of our baseline model with topic models in which 1 of the 29 questions from the baseline model was removed.

92 questions that are available across all 3 waves of the WVS used in this paper. As an additional robustness check, we include all these 92 questions in our topic model and create an extended type hierarchy. We then correlate the weights on the β positions between the original and extended models where they overlap.

Practically, this exercise allows us to ask whether the relative ordering of the original β issue-position weights changes as we add more features to the model. Note that this is more of an 'add them all in' rather than an iterative 'add one in' exercise. We adopt this approach both for the sake of brevity as well as to see how our original 4-type model is affected by a large, lateral addition of information. The concern would be that the addition of many extra features would fundamentally change the structure of the clusters and shift the ordering of the initial set of features.

Table 3.11 reports the correlations between the β-vectors from the baseline type hierarchy

and those from the extended-feature type hierarchy. Obviously, the correlation coefficients can only be calculated on the basis of the 29 original questions used in the baseline hierarchy. The correlations are very high across all the hierarchy models. Overall, we find these results to be encouraging. The same basic type structure is intact even when adding in a large amount of information. This is compatible with the idea that the extra questions/features fit in as new responses that tap into a stable set of underlying latent types.

Table 3.11: Sensitivity to Additional Features

2 Type Model					
	Type 1'	Type 2'			
Left		0.985			
Right	0.983				
3 Type Model					
	Type 1'	Type 2'	Type 3'		
Lib. Centrist		0.947			
Cons. Centrist			0.951		
Anarchist	0.923				
4 Type Model					
	Type 1'	Type 2'	Type 3'	Type 4'	
Lib. Centrist		0.944			
Cons. Centrist			0.937		
Left Anarchist	0.829				
Right Anarchist				0.631	
5 Type Model					
	Type 1'	Type 2'	Type 3'	Type 4'	Type 5'
Lib. Centrist	0.877				
Cons. Centrist			0.941		
Left Anarchist					0.800
Right Anarchist				0.970	
Market Lib.		0.987			

Notes: This table reports the correlation of the β vectors of the type hierarchy from the main paper and the type hierarchy of a topic model including all 92 consistent questions from the WVS. The prime' notation indicates the types estimated using the 92 feature topic model. We report the highest cross-model correlations for the overlapping β weights, except for the 4-type Left Anarchist case where (in the interests of exposition) we report the three highest correlations.

We stress though that both of the exercises we present here are indicative with limited formal precedents in the LDA literature. One interesting pattern here is that the Centrist types are less sensitive to changes in features relative to the Anarchist types. This fits with the intuition that the Centrist types are well-established and better defined with the Anarchist types still being more fluid. The tendency of the Anarchist types to split as we consider higher-order models (eg: 5, 6, and 7-type models) is also consistent with this assessment.

3.6.6 Appendix: Additional Type Hierarchy Information

Table 3.12: Extended Hierarchy of Types (Top Ten Features)

6 Type Model	7 Type Model
Liberal Centrist	**Liberal Centrist**
Confidence: The Civil Services	Confidence: The Civil Services
Confidence: Parliament	Confidence: Parliament
Confidence: Justice System/Courts	Confidence: Justice System/Courts
Confidence: Police	Confidence: Police
No problem Neighbours: Homosexuals	Proud of nationality
Proud of nationality	No problem Neighbours: People different race
No problem Neighbours: People different race	Not Justifiable: someone accepting a bribe
No problem Neighbours: People AIDS	No problem Neighbours: Homosexuals
No problem Neighbours: Immigrants/foreign workers	No problem Neighbours: Immigrants/foreign workers
Not Justifiable: someone accepting a bribe	Justifiable: divorce
Conservative Centrist	**Conservative Centrist**
Not Justifiable: abortion	Confidence: Police
Confidence: Police	Not Justifiable: abortion
Not Justifiable: prostitution	Confidence: Churches
Confidence: Churches	Not Justifiable: euthanasia
Confidence: Armed Forces	Confidence: Armed Forces
Not Justifiable: suicide	Not Justifiable: suicide
Not Justifiable: someone accepting a bribe	Not Justifiable: prostitution
Not Justifiable: cheating on taxes	No problem Neighbours: People different race
Not Justifiable: avoiding a fare on public transport	Not Justifiable: cheating on taxes
Not Justifiable: claiming government benefits	Not Justifiable: someone accepting a bribe
Left Anarchist	**Left Anarchist**
No Confidence: Armed Forces	No Confidence: Armed Forces
Justifiable: divorce	No Confidence: Churches
No Confidence: Churches	Justifiable: divorce
No Confidence: Major Companies	No Confidence: Major Companies
Justifiable: homosexuality	Justifiable: homosexuality
No problem Neighbours: Homosexuals	No Confidence: Parliament
Justifiable: euthanasia	No problem Neighbours: Homosexuals
No problem Neighbours: People different race	Justifiable: euthanasia
Justifiable: abortion	No problem Neighbours: People AIDS
No problem Neighbours: Immigrants/foreign workers	No problem Neighbours: People different race
Market Liberal	**Market Liberal**
No Confidence: The Press	No Confidence: Parliament
Proud of nationality	No Confidence: The Press
No Confidence: Parliament	No problem Neighbours: People different race
Confidence: Armed Forces	No problem Neighbours: Homosexuals
Confidence: Police	Confidence: Police
Not Justifiable: someone accepting a bribe	Proud of nationality
Not Justifiable: claiming government benefits	Confidence: Armed Forces
No Confidence: Labour Unions	No problem Neighbours: People AIDS
No problem Neighbours: Homosexuals	Not Justifiable: someone accepting a bribe
No problem Neighbours: People different race	No Confidence: Labour Unions
Right Anarchist ('Soft')	**Right Anarchist ('Soft')**
No Confidence: Justice System/Courts	No Confidence: Parliament
No Confidence: Armed Forces	No Confidence: Civil Services
No Confidence: Parliament	No Confidence: Justice System/Courts
No Confidence: Civil Services	No Confidence: Armed Forces
No Confidence: Police	Not Justifiable: suicide
No Confidence: Labour Unions	No Confidence: Major Companies
Not Justifiable: suicide	No Confidence: Labour Unions
No Confidence: The Press	No Confidence: The Press
No Confidence: Major Companies	No problem Neighbours: People different race
Not Justifiable: someone accepting a bribe	Not Justifiable: prostitution
Right Anarchist ('Hard')	**Right Anarchist ('Hard')**
Against Neighbours: Immigrants/foreign workers	Against Neighbours: People AIDS
Against Neighbours: People AIDS	Against Neighbours: Homosexuals
Justifiable: avoiding a fare on public transport	Against Neighbours: Immigrants/foreign workers
Against Neighbours: People different race	Against Neighbours: Drug addicts
Justifiable: cheating on taxes	If Jobs scarce: priority to (nation) people
Against Neighbours: Homosexuals	Against Neighbours: People different race
If Jobs scarce: priority to (nation) people	Not Justifiable: homosexuality
Justifiable: claiming government benefits	Proud of nationality
Against Neighbours: Drug addicts	Confidence: Armed Forces
Justifiable: euthanasia	Not Justifiable: someone accepting a bribe
	Super Anarchist ('Rage Against the Machine')
	Justifiable: avoiding a fare on public transport
	Justifiable: cheating on taxes
	Justifiable: claiming government benefits
	Justifiable: accepting a bribe
	Justifiable: euthanasia
	If Jobs scarce: priority to (nation) people
	Proud of nationality
	Justifiable: prostitution
	Justifiable: divorce
	No problem Neighbours: People different race

Notes: This table reports the 10 most important features for a n-type LDA model, where $n \in \{6, 7\}$. The types are labeled on the basis of their β-weight correlation with types in the previous level. For example, the 6-type model Liberal Centrist has a 0.96 correlation with the 5-type model Liberal Centrist.

Table 3.13: Type Hierarchy – All WVS Waves Pooled

2 Type Model	3 Type Model	4 Type Model
Left	**Liberal Centrist**	**Liberal Centrist**
No problem Neighbours: Homosexuals	Confidence: Police	Confidence: Police
No problem Neighbours: People different race	Confidence: Justice System/Courts	No problem Neighbours: Homosexuals
No problem Neighbours: People AIDS	No problem Neighbours: Homosexuals	No problem Neighbours: People AIDS
No problem Neighbours: Immigrants/foreign workers	No problem Neighbours: People different race	No problem Neighbours: People different race
Justifiable: divorce	No problem Neighbours: People AIDS	No problem Neighbours: Immigrants/foreign workers
Not Justifiable: someone accepting a bribe	No problem Neighbours: Immigrants/foreign workers	Justifiable: divorce
Justifiable: euthanasia	Proud of nationality	Not Justifiable: someone accepting a bribe
Justifiable: homosexuality	Not Justifiable: someone accepting a bribe	Proud of nationality
Not Justifiable: claiming government benefits	Confidence: The Civil Services	Not Justifiable: claiming government benefits
Proud of nationality	Not Justifiable: cheating on taxes	Not Justifiable: cheating on taxes
Right	**Conservative Centrist**	**Conservative Centrist**
Not Justifiable: someone accepting a bribe	Not Justifiable: homosexuality	Confidence: Police
Not Justifiable: suicide	Not Justifiable: abortion	Confidence: Churches
Proud of nationality	Not Justifiable: suicide	Not Justifiable: suicide
Not Justifiable: prostitution	Not Justifiable: prostitution	Proud of nationality
Not Justifiable: avoiding a fare on public transport	Proud of nationality	Confidence: Justice System/Courts
Not Justifiable: claiming government benefits	Not Justifiable: someone accepting a bribe	Not Justifiable: prostitution
Not Justifiable: cheating on taxes	Not Justifiable: avoiding a fare on public transport	Confidence: The Civil Services
Not Justifiable: abortion	Not Justifiable: claiming government benefits	Not Justifiable: someone accepting a bribe
Not Justifiable: homosexuality	Not Justifiable: cheating on taxes	Not Justifiable: abortion
No problem Neighbours: People different race	Not Justifiable: euthanasia	Confidence: Armed Forces
	Anarchist	**Left Anarchist**
	No Confidence: Civil Services	Justifiable: divorce
	No Confidence: Parliament	No Confidence: Churches
	No Confidence: Churches	No Confidence: Armed Forces
	No Confidence: Justice System/Courts	No problem Neighbours: Homosexuals
	No problem Neighbours: Homosexuals	No Confidence: Parliament
	No Confidence: Armed Forces	No Confidence: Civil Services
	No Confidence: Major Companies	No problem Neighbours: People different race
	No problem Neighbours: People different race	No problem Neighbours: People AIDS
	No problem Neighbours: People AIDS	No problem Neighbours: Immigrants/foreign workers
	No Confidence: The Press	Justifiable: euthanasia
		Right Anarchist
		No Confidence: Parliament
		No Confidence: Civil Services
		No Confidence: Labour Unions
		No Confidence: The Press
		No Confidence: Justice System/Courts
		Not Justifiable: someone accepting a bribe
		Not Justifiable: suicide
		Not Justifiable: avoiding a fare on public transport
		Not Justifiable: claiming government benefits
		Proud of nationality

Notes: This table reports the 10 most important features based on the β vectors for a n-type LDA model fitted to all waves of the WVS pooled together, where $n \in \{2, 3, 4\}$.

Table 3.14: Type Hierarchy – No Imputation

2 Type Model	3 Type Model	4 Type Model
Left	**Liberal Centrist**	**Liberal Centrist**
No problem Neighbours: Homosexuals	Confidence: Police	Confidence: Police
No problem Neighbours: People different race	Confidence: Justice System/Courts	No problem Neighbours: Homosexuals
No problem Neighbours: People AIDS	No problem Neighbours: Homosexuals	No problem Neighbours: People AIDS
No problem Neighbours: Immigrants/foreign workers	No problem Neighbours: People different race	No problem Neighbours: People different race
Justifiable: divorce	No problem Neighbours: Immigrants/foreign workers	Proud of nationality
Not Justifiable: someone accepting a bribe	No problem Neighbours: People AIDS	No problem Neighbours: Immigrants/foreign workers
Justifiable: homosexuality	Proud of nationality	Not Justifiable: someone accepting a bribe
Proud of nationality	Not Justifiable: someone accepting a bribe	Not Justifiable: claiming government benefits
Justifiable: euthanasia	Not Justifiable: cheating on taxes	Not Justifiable: cheating on taxes
Not Justifiable: claiming government benefits	Confidence: The Civil Services	Justifiable: divorce
Right	**Conservative Centrist**	**Conservative Centrist**
Not Justifiable: someone accepting a bribe	Not Justifiable: abortion	Confidence: Police
Proud of nationality	Not Justifiable: prostitution	Confidence: Churches
Not Justifiable: suicide	Not Justifiable: suicide	Confidence: Armed Forces
Not Justifiable: cheating on taxes	Proud of nationality	Not Justifiable: prostitution
Not Justifiable: prostitution	Not Justifiable: someone accepting a bribe	Not Justifiable: abortion
Not Justifiable: avoiding a fare on public transport	Not Justifiable: cheating on taxes	Not Justifiable: suicide
Not Justifiable: claiming government benefits	Not Justifiable: avoiding a fare on public transport	Proud of nationality
Not Justifiable: abortion	Not Justifiable: homosexuality	Not Justifiable: cheating on taxes
No problem Neighbours: People different race	Not Justifiable: claiming government benefits	Not Justifiable: someone accepting a bribe
Confidence: Police	Not Justifiable: euthanasia	Not Justifiable: euthanasia
	Anarchist	**Left Anarchist**
	No Confidence: Civil Services	Justifiable: divorce
	No Confidence: Parliament	Justifiable: homosexuality
	No Confidence: Churches	Justifiable: abortion
	No Confidence: Major Companies	Justifiable: euthanasia
	No Confidence: Justice System/Courts	State ownership better than private ownership
	No problem Neighbours: Homosexuals	No problem Neighbours: Homosexuals
	No Confidence: The Press	No problem Neighbours: People different race
	No problem Neighbours: People different race	No problem Neighbours: Immigrants/foreign workers
	No problem Neighbours: People AIDS	No Confidence: Churches
	No problem Neighbours: Immigrants/foreign workers	No problem Neighbours: People AIDS
		Right Anarchist
		No Confidence: Parliament
		No Confidence: Civil Services
		No Confidence: Justice System/Courts
		No Confidence: Labour Unions
		No Confidence: The Press
		No Confidence: Major Companies
		Not Justifiable: someone accepting a bribe
		Not Justifiable: claiming government benefits
		No Confidence: Police
		No problem Neighbours: People different race

Notes: This table reports the 10 most important features based on the β vectors for a n-type LDA model fitted to the subset of observation in the WVS that do not require imputation of missing values, where $n \in \{2, 3, 4\}$.

Table 3.15: Issues of Increasing Importance between Wave 2 and Wave 5

Question	Baseline	Change
Liberal Centrist		
More responsibility for government	0.004	0.222
Confidence: Armed Forces	0.430	0.199
State ownership better than private ownership	0.000	0.164
Justifiable: homosexuality	0.477	0.148
Confidence: The Civil Services	0.425	0.144
Confidence: Labour Unions	0.324	0.136
Against Neighbours: Drug addicts	0.456	0.130
No Confidence: Major Companies	0.285	0.091
Not Justifiable: prostitution	0.323	0.087
Confidence: Churches	0.328	0.075
Conservative Centrist		
No problem Neighbours: Homosexuals	0.402	0.133
No Confidence: Parliament	0.009	0.122
No problem Neighbours: People AIDS	0.445	0.095
No Confidence: Major Companies	0.150	0.088
Justifiable: homosexuality	0.000	0.072
Not Justifiable: abortion	0.595	0.065
Competition is harmful	0.206	0.053
No problem Neighbours: Drug addicts	0.278	0.052
Confidence: Armed Forces	0.618	0.045
No Confidence: The Press	0.301	0.044
Left Anarchist		
Confidence: Police	0.024	0.204
Not Justifiable: cheating on taxes	0.277	0.144
Proud of nationality	0.365	0.103
Competition is harmful	0.444	0.098
Confidence: Armed Forces	0.002	0.090
Justifiable: homosexuality	0.570	0.077
Justifiable: suicide	0.372	0.076
Justifiable: prostitution	0.463	0.076
No Confidence: The Press	0.487	0.064
If Jobs scarce: no priority to (nation) people	0.391	0.053
Right Anarchist		
Confidence: Police	0.000	0.319
Confidence: Armed Forces	0.169	0.225
Justifiable: divorce	0.045	0.150
No problem Neighbours: Homosexuals	0.372	0.149
Justifiable: homosexuality	0.000	0.147
Justifiable: euthanasia	0.058	0.135
No Confidence: Churches	0.378	0.094
Against Neighbours: Immigrants/foreign workers	0.225	0.086
No problem Neighbours: People AIDS	0.431	0.077
More responsibility for government	0.258	0.062

Notes: This table reports the 10 feature of each type which show the biggest increase in weight from wave 2 to wave 5. Column 2 reports the baseline value in wave 2 and column 3 reports the change from wave 2 to wave 5.

Table 3.16: Issue Position Differences between Types (4 Type Model)

	Difference Lib. Centrist	Difference Cons. Centrist	Difference Left Anarchist	Difference Right Anarchist
L. Centrist		Justifiable: divorce Justifiable: euthanasia Justifiable: homosexuality Justifiable: abortion No Confidence: Churches No problem Neighbours: Homosexuals If Jobs scarce: no priority to natives No problem Neighbours: People AIDS Justifiable: prostitution Competition is good	Confidence: Justice System/Courts Confidence: Police Confidence: Armed Forces Confidence: The Civil Services Confidence: Parliament Competition is good Against Neighbours: Drug addicts Confidence: Major Companies More responsibility for people Not Justifiable: avoiding fare on pub. trans.	Confidence: Justice System/Courts Justifiable: divorce Confidence: The Civil Services Justifiable: abortion Justifiable: homosexuality Confidence: Police Justifiable: euthanasia Confidence: Parliament Confidence: Major Companies Confidence: Labour Unions
C. Centrist	Not Justifiable: abortion Not Justifiable: euthanasia Not Justifiable: homosexuality Not Justifiable: divorce Confidence: Churches Not Justifiable: prostitution Not Justifiable: suicide If Jobs scarce: priority to natives Against Neighbours: People AIDS Against Neighbours: Homosexuals		Confidence: Churches Not Justifiable: abortion Not Justifiable: prostitution Not Justifiable: euthanasia Confidence: Armed Forces Confidence: The Civil Services Confidence: Justice System/Courts Confidence: Police Not Justifiable: suicide Confidence: Parliament	Confidence: Justice System/Courts Confidence: The Civil Services Confidence: Parliament Confidence: Major Companies Confidence: Police Confidence: Churches Confidence: Labour Unions Confidence: Press Confidence: Armed Forces Not Justifiable: euthanasia
L. Anarchist	No Confidence: Armed Forces No Confidence: Justice System/Courts No Confidence: Civil Services Competition is harmful No Confidence: Parliament No Confidence: Police No Confidence: Churches No problem Neighbours: Drug addicts Justifiable: avoiding fare on pub. trans. More responsibility for government	No Confidence: Churches Justifiable: divorce No Confidence: Armed Forces Justifiable: euthanasia No Confidence: Civil Services Justifiable: homosexuality Justifiable: abortion No Confidence: Parliament No Confidence: Justice System/Courts Justifiable: prostitution		Justifiable: abortion Justifiable: divorce Justifiable: homosexuality Justifiable: euthanasia Justifiable: prostitution Competition is harmful No problem Neighbours: Drug addicts Justifiable: suicide Justifiable: avoiding fare on pub. trans. No Confidence: Churches
R. Anarchist	No Confidence: Justice System/Courts No Confidence: Civil Services No Confidence: Parliament Not Justifiable: abortion No Confidence: Labour Unions No Confidence: The Press No Confidence: Major Companies Not Justifiable: homosexuality No Confidence: Police Not Justifiable: euthanasia	No Confidence: Civil Services No Confidence: Justice System/Courts No Confidence: Parliament No Confidence: Major Companies No Confidence: Labour Unions No Confidence: The Press No Confidence: Churches No Confidence: Police No Confidence: Armed Forces Against Neighbours: Drug addicts	Not Justifiable: prostitution Not Justifiable: suicide Not Justifiable: abortion Against Neighbours: Drug addicts If Jobs scarce: priority to natives Not Justifiable: homosexuality Not Justifiable: avoiding fare on pub. trans. Competition is good Not Justifiable: euthanasia No Confidence: Labour Unions	

Notes: This table reports the 10 features for which there exists the largest differences between the 4 ideological types created by LDA. The model is fitted to the 5th wave of the sample. The type labels are chosen by the author.

3.6.7 Appendix: Cross-Check of Results with European Social Study

The European Social Study(ESS) is a biannual survey of 37 European countries covering the years from 2002 until 2016. For our replication exercise, we use all countries in the ESS that also appear in the WVS and were used in our main analysis. Overall, 13 of our original 17 countries also appear in the ESS (Germany, Great Britain, France, Denmark, Spain, Finland, Portugal, Austria, Belgium, Italy, Ireland, Netherlands, Iceland). Similarly, we create a subsample of the ESS waves that aligns with the waves of the WVS. We use ESS rounds 1 to 5 (2002 - 2010) which are comparable to the 4th and 5th wave of the WVS (Wave 4: 1999-2004 and Wave 5: 2005-2009). We further select a set questions from the ESS that cover similar issues to those we used from the WVS. Table 3.17 provides an overview over the ESS questions as well as their scale. Identical to the main results of the paper, we recode questions into 2 binary features indicating support and opposition to issues .

As the next step, we fit LDA models with an increasing number of types to the ESS data to produce a type hierarchy. Identical to the results in Table 3.2, we report the 'top ten' features for each ESS type in Table 3.18. It should be apparent that since we use a different set of questions the ESS types can never be identical to the WVS types. What is important for our purpose is that the resulting types recover a similar ideological spectrum.

For the basic 2-type model in the first column, the two types are distinguished by their trust in institutions. While the first type, which we label as 'Centrist' trusts the police, the legal system and is satisfied with the democracy in the country, the second type (labelled as 'Anarchist') does not trust politicians and political parties and is unsatisfied with the national government. Interestingly, this shows that in the ESS data the 'Anarchist' type already arises in the 2 type model.

The second column reports the top features for the 3-type model. The 'Centrist' type remains more or less unchanged, but we observe a split of the 'Anarchist' type along immigration issues. On the one hand, the 'Left Anarchist' supports immigration and gay rights. Moreover, this type considers it important to take care of people and treat them equally. The 'Right Anarchist' on the other hand opposes immigration and puts a larger weight on security and safety.

In the third column, we show the top features for the 4-type model. In the 4-type model, a new split between two 'Centrist' types emerges. One important difference between the two

Table 3.17: Selected Question from the ESS

ESS Variable Code	Question	Scale
ppltrst	Most people can be trusted or you can't be too careful	
trstprl	Trust in country's parliament	
trstlgl	Trust in the legal system	
trstplc	Trust in the police	
trstplt	Trust in politicians	
trstprt	Trust in political parties	
trstep	Trust in the European Parliament	
trstun	Trust in the United Nations	
stflife	How satisfied with life as a whole	
stfeco	How satisfied with present state of economy in country	
stfgov	How satisfied with the national government	
stfdem	How satisfied with the way democracy works in country	
stfedu	State of education in country nowadays	
stfhlth	State of health services in country nowadays	
gincdif	Government should reduce differences in income levels	
freehms	Gays and lesbians free to live life as they wish	
imsmetn	Allow many/few immigrants of same race/ethnic group as majority	
imdfetn	Allow many/few immigrants of different race/ethnic group from majority	
impcntr	Allow many/few immigrants from poorer countries outside Europe	
imbgeco	Immigration bad or good for country's economy	
imueclt	Country's cultural life undermined or enriched by immigrants	
imwbcnt	Immigrants make country worse or better place to live	
rlgdgr	How religious are you	
rlgatnd	How often attend religious services apart from special occasions	
pray	How often pray apart from at religious services	
ipeqopt	Important that people are treated equally and have equal opportunities	
impsafe	Important to live in secure and safe surroundings	
ipfrule	Important to do what is told and follow rules	
ipudrst	Important to understand different people	
ipgdtim	Important to have a good time	
impfree	Important to make own decisions and be free	
iphlppl	Important to help people and care for others well-being	
ipstrgv	Important that government is strong and ensures safety	
ipbhprp	Important to behave properly	
iprspot	Important to get respect from others	
iplylfr	Important to be loyal to friends and devote to people close	
impenv	Important to care for nature and environment	
imptrad	Important to follow traditions and customs	

Notes: This table reports the questions selected from the European Social Study.

'Centrist' types is the importance of religions, traditions and customs. Further, the two types differ based on the importance they attribute to safety, but both profess trust in the legal system.

Overall, the type structure that emerges from the ESS is reasonably similar to the types that emerge in the WVS. We again find that types split apart based on their trust in institutions. This allows us to label types as 'Centrist' and 'Anarchist'. Additionally, we observe type characteristics that are broadly in line with the left-right spectrum. For example, one of the important dividing issues is immigration. The social issues which define the types differ across the two datasets but this is mainly a result of the differences in the question set. The ESS simply does not contain questions concerning support and opposition abortion and

suicide, neither does the WVS contain a detailed set of questions concerning immigration. We, therefore, view this exercise as useful corroboration that our core finding of ideological types that are differentiated by trust in institutions holds across independent datasets.

Table 3.18: Type Hierarchy as Created with ESS Data

2 Type Model	3 Type Model	4 Type Model
Centrist	**Centrist**	**Conservative Centrist**
Satisfied with life as a whole	Satisfied: Democracy in country	Satisfied: Democracy in country
Trust in the police	Trust in the police	Trust in the police
Important: To be loyal to friends	Satisfied with life as a whole	Important to behave properly
Satisfied: Democracy in country	Trust in the legal system	Satisfied with life as a whole
Important: That people are treated equally	Trust in country's parliament	Important to follow traditions and customs
Important: To help people and care for others	Satisfied with state of education	Trust in the legal system
Important: To care for environment	Important: To be loyal to friends	Important: To help people and care for others
Important to understand different people	Trust in the United Nations	Important: To care for environment
Trust in the legal system	Satisfied with state of health services	Important: To be loyal to friends
Important to make own decisions and be free	Important: To care for environment	Important: That government ensures safety
Anarchist	**Left Anarchist**	**Left Anarchist**
No Trust in politicians	Allow immigrants of same race/ethnic group	Allow immigrants of same race/ethnic group
No Trust in political parties	Allow immigrants of different race/ethnic group	No Trust in politicians
Important: To be loyal to friends	Allow immigrants from poorer countries	Allow immigrants of different race/ethnic group
Important: That people are treated equally	No Trust in politicians	Allow immigrants from poorer countries
Not satisfied with the national government	Important: That people are treated equally	No Trust in political parties
Important: To help people and care for others	Important: To be loyal to friends	Important: That people are treated equally
Important: To care for environment	Gays and lesbians free to live life as they wish	Important: To help people and care for others
Important to make own decisions and be free	Important: To help people and care for others	Important: To be loyal to friends
Important to live in secure and safe surroundings	Important to understand different people	Important: To care for environment
Important: That government ensures safety	Important to make own decisions and be free	Important to understand different people
	Right Anarchist	**Right Anarchist**
	Not Allow immigrants of different race/ethnic group	Not Allow immigrants of different race/ethnic group
	Not Allow immigrants from poorer countries	Not Allow immigrants from poorer countries
	Immigrants make country worse place to live	Not Allow immigrants of same race/ethnic group
	Not Allow immigrants of same race/ethnic group	Immigrants make country worse place to live
	Important: To be loyal to friends	No Trust in politicians
	No Trust in politicians	Immigration bad for country's economy
	Important to live in secure and safe surroundings	No Trust in political parties
	Important: That government ensures safety	Important: To be loyal to friends
	Immigration bad for country's economy	Important to live in secure and safe surroundings
	No Trust in political parties	Important: That government ensures safety
		Liberal Centrist
		Not Important to follow traditions and customs
		Not Important to do what is told and follow rules
		Not often pray apart from at religious services
		Not often attend religious services
		Not Important to behave properly
		Not religious
		Gays and lesbians free to live life as they wish
		Not Important to live in secure and safe surroundings
		Not Important: That government ensures safety
		Trust in the legal system

Notes: This table reports the 10 most important features for a n-type LDA model fit to the European Social Study, where $n \in \{2, 3, 4, \}$.

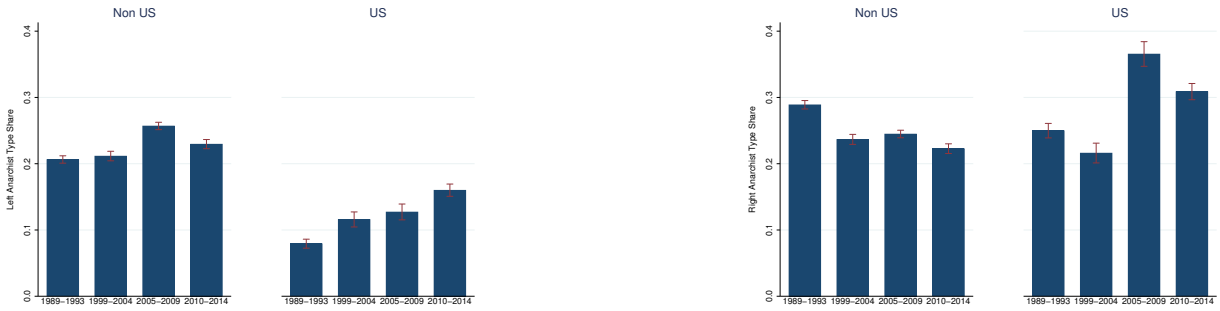
3.6.8 Appendix: Robustness 6th Wave of the WVS

Our main analysis is based on the 2nd, 3rd and 5th wave of the World Value Survey (WVS) and European Value Study (EVS). The combination of these two surveys significantly increases the number of European countries that are covered in our data. For this reason, we excluded the 4th and 6th wave of the WVS as there are no corresponding waves of the EVS.[94] We can extend our analysis past 2010 by making use of the 6th wave of the WVS but in this process, the sample of available countries shrinks to Germany, Netherlands, Spain, and the US. Due to the shrinking and changing sample, we regard the following analysis as solely suggestive.

Fitting a 4-type LDA topic model to the 6th wave of the WVS leads to broadly similar types. We again observe a liberal centrist and conservative centrist type in the data (correlations of 0.94 and 0.78 respectively with their wave 5 equivalent). Also, the right anarchist type emerges in the LDA model (correlation of 0.72 with wave 5 equivalent). Only the left anarchist type is less clearly apparent (correlation 0.47 with wave 5 equivalent). This result is most likely driven by the fact that many countries with large left anarchist type shares are not any longer in the wave 6 sample.

Figure 3.11: Type Shares - US vs non-US (Wave 6)

(a) Left Anarchist Types **(b) Right Anarchist Types**



Notes: This figure compares the levels of θ type shares across waves for the Left Anarchist and Right Anarchist types. We pool all 3 non-US countries (Germany, Netherlands, Spain) and contrast them to the US. The pooling for the non-US sample is based on WVS sample weights. The timing of the waves is Wave 2 (1989-1993), Wave 4 (1999-2004) and Wave 5 (2005-2009). 95% confidence intervals are reported in red.

To circumvent the problem that the shifting type shares might render our wave 6 analysis less meaningful, we keep the β-type vectors from wave 5 constant and only use LDA to generate new individual-level type shares α for wave 6. Based on the resulting type shares we analyze if any further changes in the type composition occurred in wave 6. In particular,

[94] A 7th wave of the WVS is currently in progress and should be completed in 2020. For this 7th wave of the WVS there will be a corresponding wave of the EVS.

we reproduce Figure 3.6 based on the countries that are available in all 4 waves. The results are presented in Figure 3.11.

Overall, our findings are similar when we include the 6th wave. The anarchist types stabilized at a its high level in the US. We only observe a slight shift from the right anarchist to the left anarchist type. For the other countries in the sample, we do not observe any major shifts in the prevalence of the anarchist types. If anything the anarchist type shares appear to decrease slightly.

3.6.9 Appendix: Additional Details on Populist Parties

A list of European parties that can be classified as populists in 2019 was prepared by Rooduijn et al. (2019). Their classification is based on the following definition:

> "Populist parties: parties that endorse the set of ideas that society is ultimately separated into two homogeneous and antagonistic groups, "the pure people" versus "the corrupt elite," and which argues that politics should be an expression of the volonté générale (general will) of the people (Mudde, 2004)."

As the list does not contain any information for parties outside of Europe, we further code the Reform Party in the US as populist parties based on the (see http://www.reformparty.org/). Lastly also the NDP in Canada is classified as populist as it exhibited populists tendencies during our observation period (see https://www.thecanadianencyclopedia.ca/en/article/populism).

To achieve a consistent coding of parties across waves, we also classify predecessor parties as populist. For example, the German party "Die Linke" is listed in Rooduijn et al. (2019). Hence, we also code the party "Partei des demokratischen Sozialismus" as populist.

3.6.10 Appendix: Additional Details on the Polarisation Measure

The Esteban and Ray (1994) measure of polarisation is based on three axioms. These three axioms aim to capture sensible assumptions about how own-group identification and out-group alienation contribute to an overall index of polarisation.

Figure 3.12 illustrates the three axioms of Esteban and Ray (1994) graphically. The first axiom states that polarisation increases if two small masses b and c that are close to each other are joined at their midpoint (see panel (a) of Figure 3.12). The intuition behind

Table 3.19: List of Populist Parties

Country	Party
Austria	FPÖ
Austria	Alliance for the Future of Austria
Austria	Dr. Martin's List - For Democracy
Belgium	Front National
Belgium	Vlaams Blok
Belgium	Vlaams Belang
Canada	NDP
Denmark	Danish People Party
Denmark	Progress Party
Finland	True Finns
France	Front National
France	Le Front National de Jean-Marie le Pen
France	Le Front National de Bruno Megret
Germany	Partei des demokratischen Sozialismus
Iceland	Citizen Movement
Ireland	Sinn Fein
Italy	Forza Italia
Italy	Northern League
Netherland	Party for Freedom
Netherland	Socialistische Partij
United Kingdom	UK Independence Part
United Kingdom	Sinn Fein
United States	Reform Party

Notes: This table reports the parties that were coded as populist based on the information from Rooduijn et al. (2019)

this axiom is that the joining of the masses increases the own-group identification of the now joined smaller masses, while the average distance and out-group alienation with respect to other major societal group a stay unchanged.

The second axiom states that polarisation increases if a small mass of people b moves closer to the side of the spectrum where fewer people are concentrated (see panel (b) of Figure 3.12). Put simply, this change increases polarisation because while the mass b has moved closer to group c it has also moved further away from another group a. Since mass a is larger than mass c, the overall alienation effect increases.

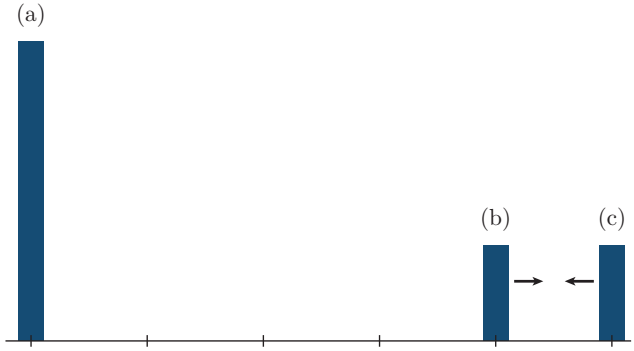
The third axiom states that polarisation increases if mass is shifted equally from a central mass b to two lateral masses a and c that are each equally far away from the central mass (see panel (c) of Figure 3.12). This axiom captures the effect of the disappearing centre. If mass shifts equally from the centre to the fringes of the spectrum the own-group identification at the fringes increases while the overall out-group alienation increases as well.

Esteban and Ray (1994) prove that any measure of polarisation that fulfills these three axioms must be of the form:

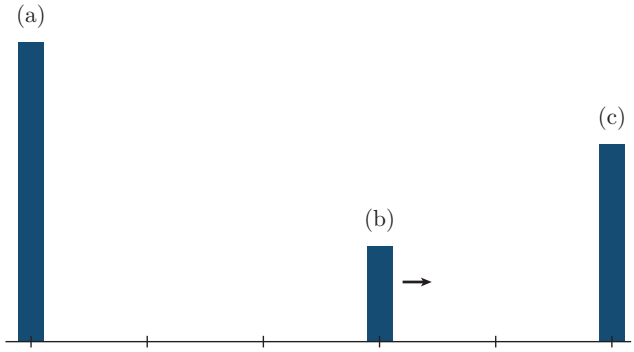
$$P(\pi, y) = \kappa \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i^{1+\nu} \pi_j |y_i - y_j| \tag{3.18}$$

Figure 3.12: Axioms of Esteban & Ray 1994

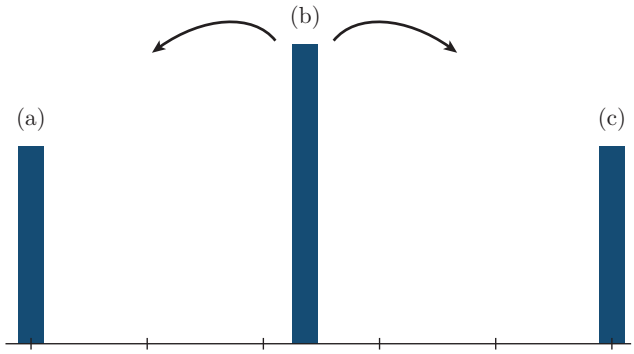
(a) Axiom 1



(b) Axiom 2



(c) Axiom 3



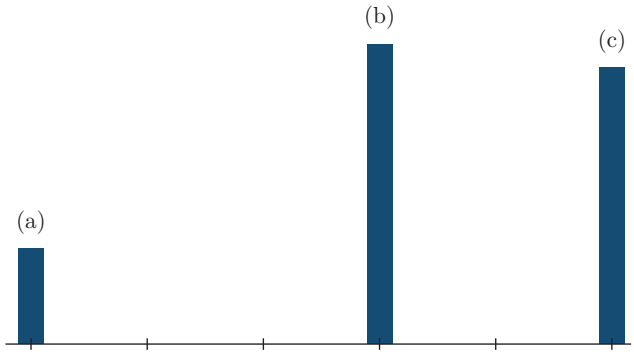
Notes: This figure illustrates the 3 main axioms use in Esteban and Ray (1994) to derive the polarisation measure.

The axioms hold for values of $\nu \in [0, 1.6]$. The sensitivity parameter ν also influences the maximal possible value of the polarisation measure. Note that the measure will not be bounded between [0,1]. Esteban and Ray (1994) suggest a potential fourth axiom that would make it possible to narrow the possible interval of $\nu \in [1, 1.6]$.

This fourth axiom is illustrated in Figure 3.13. The axiom states that moving mass from a small mass a to a larger mass c will increase polarisation. Hence, the axiom makes an assumption on the importance of small groups within a society. On the one hand, moving mass from a to c reduced the distance between the groups and therefore lowered polarisation. On the other hand the mass a is small in comparison to b and c and hence the effect of group a for overall polarisation might be negligible, while increasing the mass of c can increase societal tension.

The polarisation sensitivity parameter ν here captures the relative sizes of a and c for which polarisation will increase. The larger is ν the smaller is the importance of a for overall polarisation. It is a priori not clear whether this axiom is sensible in our context. Hence, we do not restrict the range of polarisation sensitivity to $\nu > 1$.

Figure 3.13: Additional Axiom of Esteban & Ray 1994



Notes: This figure illustrates the 4th axiom suggested in Esteban and Ray (1994). This axioms is not necessary to derive the form of the polarisation measure but it allows for restrictions to the possible range of ν.

Extending the Esteban and Ray (1994) Measure to Higher Dimension

The Esteban and Ray (1994) measure was originally constructed for one-dimensional indicators (e.g. the income distribution). Our measure extends the measure to the four dimensions of our ideological type space. We assume that an individual identifies with groups based on his or her dominant type share, since in our model the four ideological types are the most natural line for group delineations.

Theoretically, it would also be possible to define groups based on discrete intervals of the type share distribution, such that a type would be defined by a specific interval in the four-dimensional ideological type space (e.g. [0,0.1] Liberal Centrist, [0.2,0.3] Conservative Centrist, [0.4,0.5] Left and Right Anarchist). This would lead to a far greater number of ideological groups. The problem with this approach is that it is not obvious to decide on an interval length such that we can plausibly assume sufficient degrees of separation between these groups.

If the groups are defined by the dominant type share of each individual, intuitively, the alienation between these groups will be based on differences in type shares. The only alteration to the original measure then is the fact that in our case the groups can differ along four dimensions rather than a single variable y. We hence define the overall out-group alienation as the sum of the type share differences between different groups.

Robustness Esteban-Ray Measure

So far we have not addressed the question of the choice of ν. As explained above any $\nu \in [0, 1.6]$ leads to a measure of polarisation that fulfils the axioms of Esteban and Ray (1994). As a robustness exercise, we calculate the Esteban-Ray measure for several values of ν. Table 3.20 reports the ranking our countries by their polarisation over the three waves conditional on the choice of ν. It is important to note that the values of the polarisation measure are not comparable across different ν, since dependent on ν the maximal possible polarisation level varies.

Our main finding for the rising level of polarisation in the US holds for all except the largest values of ν. As long as $\nu < 1$ the US emerges as the most polarised country in our sample. The results for $\nu = 1.6$ differ, since for high values of ν the importance of small groups in society is diminished. Hence, in this case, the polarisation P measure for the US - where we observe four comparably sized ideological groups - is lower than for other values of ν. In contrast, measured polarisation is higher in countries with one large ideological group, e.g. the Conservative Centrist in Malta or Liberal Centrist in Denmark.

Overall, the results seem to point towards the fact that values of $\nu < 1$ lead to a more balanced polarisation ranking across countries. The fact that for $\nu = 1.6$ countries such as Denmark, Iceland, Finland and Canada - all of which are usually considered harmonious

societies - end up at top of the ranking seems counterintuitive. Based on these findings we set $\nu = 0.5$ as the baseline value for polarisation sensitivity in our main P measure.

Table 3.20: Esteban-Ray Polarisation Measure for different ν

Panel A: Wave 2

$\alpha = 0$		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 1.6$	
Country	Pol. Measure	Country	Pol. Measure	Country	Pol. Measure	Country	Pol. Measure
Spain	1.077	Spain	0.555	Malta	0.356	Malta	0.226
France	1.059	Austria	0.539	North Ireland	0.315	North Ireland	0.178
Belgium	1.058	France	0.539	Portugal	0.308	Portugal	0.177
Italy	1.024	Belgium	0.538	Austria	0.301	Ireland	0.167
Netherlands	1.024	Malta	0.532	Netherlands	0.296	Netherlands	0.161
Germany	1.017	Netherlands	0.531	Spain	0.293	Austria	0.154
Austria	1.006	North Ireland	0.530	United States	0.287	United States	0.150
Great Britain	0.990	Italy	0.528	Ireland	0.285	Canada	0.143
North Ireland	0.958	Germany	0.519	Italy	0.282	Denmark	0.142
Canada	0.954	Great Britain	0.518	France	0.281	Spain	0.139
Finland	0.929	Portugal	0.508	Belgium	0.279	Iceland	0.138
United States	0.921	United States	0.504	Great Britain	0.278	Italy	0.138
Iceland	0.902	Canada	0.503	Canada	0.278	Great Britain	0.135
Portugal	0.898	Finland	0.478	Germany	0.271	France	0.132
Ireland	0.853	Iceland	0.478	Iceland	0.266	Finland	0.131
Malta	0.849	Ireland	0.472	Finland	0.258	Belgium	0.130
Denmark	0.827	Denmark	0.442	Denmark	0.255	Germany	0.128

Panel A: Wave 4

$\alpha = 0$		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 1.6$	
Country	Pol. Measure	Country	Ray Measure	Country	Pol. Measure	Country	Pol. Measure
Spain	1.151	Spain	0.576	Malta	0.349	Malta	0.230
Austria	1.070	Austria	0.553	North Ireland	0.300	Iceland	0.188
France	1.066	Great Britain	0.546	Ireland	0.297	Denmark	0.182
Belgium	1.057	France	0.540	United States	0.295	Netherlands	0.163
Germany	1.053	Germany	0.537	Austria	0.294	Ireland	0.162
Great Britain	1.052	Italy	0.537	Canada	0.291	Canada	0.161
Italy	1.038	United States	0.533	Great Britain	0.291	North Ireland	0.157
United States	1.005	North Ireland	0.530	Spain	0.289	Finland	0.157
North Ireland	0.987	Belgium	0.529	Italy	0.287	United States	0.151
Canada	0.952	Ireland	0.514	Finland	0.286	Portugal	0.150
Portugal	0.946	Malta	0.509	Netherlands	0.284	Austria	0.141
Ireland	0.945	Canada	0.507	France	0.281	Great Britain	0.140
Finland	0.935	Finland	0.499	Germany	0.280	Italy	0.139
Netherlands	0.918	Portugal	0.498	Portugal	0.279	France	0.133
Malta	0.794	Netherlands	0.487	Iceland	0.279	Germany	0.131
Iceland	0.755	Iceland	0.428	Belgium	0.266	Spain	0.126
Denmark	0.669	Denmark	0.376	Denmark	0.254	Belgium	0.117

Panel A: Wave 5

$\alpha = 0$		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 1.6$	
Country	Pol. Measure	Country	Pol. Measure	Country	Pol. Measure	Country	Pol. Measure
United States	1.068	United States	0.563	Malta	0.320	Malta	0.208
Netherlands	1.063	Netherlands	0.543	United States	0.306	Denmark	0.175
Austria	1.057	Austria	0.534	Canada	0.291	Iceland	0.168
Spain	1.054	Spain	0.530	North Ireland	0.291	Finland	0.162
Germany	1.032	Canada	0.528	Ireland	0.291	North Ireland	0.157
France	1.020	Ireland	0.523	Netherlands	0.285	United States	0.152
Belgium	1.000	Great Britain	0.520	Finland	0.284	Ireland	0.150
Canada	0.999	Germany	0.519	Portugal	0.283	Canada	0.149
Great Britain	0.999	France	0.518	Great Britain	0.280	Portugal	0.147
Ireland	0.985	North Ireland	0.510	Austria	0.275	Great Britain	0.138
Italy	0.954	Belgium	0.507	Italy	0.273	Italy	0.138
North Ireland	0.949	Portugal	0.502	France	0.270	Netherlands	0.137
Portugal	0.930	Italy	0.499	Spain	0.269	France	0.128
Finland	0.916	Finland	0.489	Germany	0.264	Austria	0.126
Iceland	0.767	Malta	0.475	Belgium	0.263	Belgium	0.124
Malta	0.756	Iceland	0.413	Iceland	0.258	Spain	0.120
Denmark	0.595	Denmark	0.334	Denmark	0.233	Germany	0.118

Notes: This table reports the polarisation measure for different ν. For more details see text.

3.6.11 Appendix: Comparison of LDA to PCA, Factor Analysis and k-means

This section provides a comparison between Latent Dirichlet Allocation (LDA) and the other alternative machine learning dimensionality reduction techniques, specifically Principal Component Analysis (PCA), Factor Analysis (FA) and k-means clustering. At their core, all of these techniques aim to reduce high dimensional data to a set of more easily interpretable topics, components, factors or clusters. Differences arise in the way these lower-dimensional representations of the data are constructed.

As we have outlined in detail in the main part of the paper, LDA relies on a generative model that makes assumptions about the data generating process and allows for a direct interpretation of the latent objects as topics. Furthermore, the LDA model was specifically designed for the analysis of sparse multinomial data.

PCA, on the other hand, relies on a truncated singular value decomposition to derive components that explain the maximum possible amount of variance in the data while keeping all components orthogonal to each other. The truncated singular value decomposition is based on decomposing the original $O \times F$ data matrix D of rank R with O observation and F features into three matrices such that $D = U\Sigma W^T$, where U is a $O \times R$ orthogonal matrix, W^T is a $R \times F$ orthogonal matrix, and Σ is a $R \times R$ diagonal matrix. Afterwards, PCA truncates the resulting matrices by removing the rows and columns associated with the smallest eigenvalues in the matrix Σ. This truncation process reduces the dimensions of the matrices to a user-chosen number of components C, such that U becomes U_C of dimension $O \times C$, Σ becomes Σ_C of dimension $C \times C$, and W^T becomes W_C^T of dimension $C \times F$.

Each of the resulting components are orthogonal to each other and represent a linear combination of the original data weighted by eigenvectors. This highlights two important limitations of PCA for our application. Neither is it obvious that the ideological types (components) we want to find in the data should be orthogonal to each other nor are they necessarily a linear combination of the data. As a result, the ideological type hierarchy created by PCA (see Table 3.21) is less coherent than the types created by LDA.

Similar problems arise when using FA. FA represents the original data as a linear combination of factors such that $D = C + \beta \cdot F + \epsilon$, where D is the original data matrix, C is a vector of constants F is the factor matrix, β are the factor loadings and ϵ a vector of Gaussian noise. The advantage of FA in comparison to PCA is that it accounts for random measurement

error through the ϵ vector and hence allows for heteroscedastic noise. Nevertheless, FA still uses a linear model to decompose the data. Due to the linear model the ideological type generate by FA (see Table 3.22) are less coherent than the LDA results. Note that the change in type 1 and 2 from the 2 type to the 3 type model is driven by the change in the signs of the factor loadings. The factors still load on the same features, they just point in the opposite direction.

Last, k-means is a clustering algorithm that minimizes the distance of the original data to a user-chosen number of centroids. As any other clustering algorithm, k-means assigns each observation to a unique cluster. This seems counterintuitive in our case since people do not necessarily subscribe to a single political ideology. For example, people might be liberal when it comes to social issues but conservative with regard to economic questions. While LDA captures this its mixture of ideological types, k-means cannot account for this.[95] Moreover, as discussed by Ding and He (2004) k-means clustering represents a discrete cluster solution to the components derived by PCA. As such k-means suffers from similar shortcomings as PCA and the derived ideological types (see Table 3.23) also are less coherent in comparison to LDA.

[95]PCA and FA also allow for 'mixed membership' through different component and factor loadings.

Table 3.21: Hierarchy of Types (Top Ten Features) as created by PCA

2 Type Model	3 Type Model	4 Type Model
Type 1	**Type 1**	**Type 1**
No Confidence: Churches	No Confidence: Churches	No Confidence: Churches
No Confidence: Civil Services	No Confidence: Civil Services	No Confidence: Civil Services
No Confidence: Parliament	No Confidence: Parliament	No Confidence: Parliament
No Confidence: Armed Forces	No Confidence: Armed Forces	No Confidence: Armed Forces
No Confidence: Justice System/Courts	No Confidence: Justice System/Courts	No Confidence: Justice System/Courts
No Confidence: Police	No Confidence: Police	No Confidence: Police
No Confidence: Major Companies	No Confidence: Major Companies	No Confidence: Major Companies
Justifiable: euthanasia	Justifiable: euthanasia	Justifiable: euthanasia
Justifiable: abortion	Justifiable: abortion	Justifiable: abortion
Justifiable: divorce	Justifiable: divorce	Justifiable: divorce
Type 2	**Type 2**	**Type 2**
Not Justifiable: abortion	Not Justifiable: abortion	Not Justifiable: abortion
Not Justifiable: homosexuality	Not Justifiable: homosexuality	Not Justifiable: homosexuality
Not Justifiable: euthanasia	Not Justifiable: euthanasia	Not Justifiable: euthanasia
No Confidence: Justice System/Courts	No Confidence: Justice System/Courts	No Confidence: Justice System/Courts
No Confidence: Parliament	No Confidence: Parliament	No Confidence: Parliament
No Confidence: Civil Services	No Confidence: Civil Services	No Confidence: Civil Services
Not Justifiable: divorce	Not Justifiable: divorce	Not Justifiable: divorce
No Confidence: Labour Unions	No Confidence: Labour Unions	No Confidence: Labour Unions
Not Justifiable: prostitution	Not Justifiable: prostitution	Not Justifiable: prostitution
No Confidence: The Press	No Confidence: The Press	No Confidence: The Press
	Type 3	**Type 3**
	More responsibility for people	More responsibility for people
	Against Neighbours: Drug addicts	Against Neighbours: Drug addicts
	Competition is good	Competition is good
	Private better than state ownership	Private better than state ownership
	If Jobs scarce: priority to (nation) people	If Jobs scarce: priority to (nation) people
	Justifiable: euthanasia	Justifiable: euthanasia
	Against Neighbours: People AIDS	Against Neighbours: People AIDS
	Confidence: Armed Forces	Confidence: Armed Forces
	No Confidence: Labour Unions	No Confidence: Labour Unions
	Against Neighbours: Immigrants/foreign workers	Against Neighbours: Immigrants/foreign workers
		Type 4
		More responsibility for government
		Against Neighbours: People AIDS
		Confidence: Labour Unions
		If Jobs scarce: priority to (nation) people
		Against Neighbours: Homosexuals
		Against Neighbours: Immigrants/foreign workers
		Confidence: Press
		Competition is harmful
		State ownership better than private ownership
		Against Neighbours: Drug addicts

Notes: This table reports the 10 most important features for a n-type Principal Component Analysis model, where $n \in \{2, 3, 4\}$.

Table 3.22: Hierarchy of Types (Top Ten Features) as created by FA

2 Type Model	3 Type Model	4 Type Model
Type 1	**Type 1**	**Type 1**
Confidence: Churches	No Confidence: Civil Services	No Confidence: Civil Services
Confidence: The Civil Services	No Confidence: Churches	No Confidence: Churches
Confidence: Parliament	No Confidence: Parliament	No Confidence: Parliament
Not Justifiable: abortion	No Confidence: Justice System/Courts	No Confidence: Justice System/Courts
Confidence: Armed Forces	No Confidence: Armed Forces	Justifiable: abortion
Confidence: Justice System/Courts	No Confidence: Police	No Confidence: Police
Confidence: Police	Justifiable: abortion	No Confidence: Armed Forces
Not Justifiable: prostitution	Justifiable: euthanasia	Justifiable: euthanasia
Not Justifiable: euthanasia	Justifiable: divorce	Justifiable: divorce
Confidence: Major Companies	No Confidence: Major Companies	Justifiable: homosexuality
Type 2	**Type 2**	**Type 2**
Justifiable: abortion	Not Justifiable: abortion	Not Justifiable: abortion
Justifiable: homosexuality	No Confidence: Parliament	No Confidence: Parliament
Confidence: Parliament	No Confidence: Civil Services	No Confidence: Civil Services
Confidence: The Civil Services	Not Justifiable: homosexuality	Not Justifiable: homosexuality
Justifiable: divorce	Not Justifiable: euthanasia	No Confidence: Justice System/Courts
Confidence: Justice System/Courts	No Confidence: Justice System/Courts	Not Justifiable: euthanasia
Justifiable: euthanasia	Not Justifiable: divorce	Not Justifiable: divorce
Confidence: Labour Unions	Not Justifiable: prostitution	No Confidence: Labour Unions
Confidence: Police	No Confidence: Labour Unions	Not Justifiable: prostitution
Confidence: Press	Not Justifiable: suicide	No Confidence: The Press
	Type 3	**Type 3**
	Against Neighbours: Immigrants/foreign workers	Against Neighbours: Immigrants/foreign workers
	Against Neighbours: People different race	Against Neighbours: People different race
	Against Neighbours: People AIDS	Against Neighbours: People AIDS
	Against Neighbours: Homosexuals	Against Neighbours: Homosexuals
	If Jobs scarce: priority to (nation) people	If Jobs scarce: priority to (nation) people
	Not Justifiable: homosexuality	Not Justifiable: homosexuality
	Against Neighbours: Drug addicts	Against Neighbours: Drug addicts
	Not Justifiable: abortion	Not Justifiable: abortion
	Not Justifiable: divorce	Not Justifiable: divorce
	No Confidence: Justice System/Courts	No Confidence: Justice System/Courts
		Type 4
		Not Justifiable: cheating on taxes
		Not Justifiable: claiming government benefits
		Not Justifiable: avoiding a fare on public transport
		Not Justifiable: someone accepting a bribe
		Justifiable: homosexuality
		No problem Neighbours: Homosexuals
		No problem Neighbours: People AIDS
		Justifiable: divorce
		No Confidence: Major Companies
		Competition is good

Notes: This table reports the 10 most important features for a n-type Factor Analysis model, where $n \in \{2, 3, 4\}$.

Table 3.23: Hierarchy of Types (Top Ten Features) as created by k-means

2 Type Model	3 Type Model	4 Type Model
Type 1	**Type 1**	**Type 1**
Not Justifiable: someone accepting a bribe	No problem Neighbours: People different race	Proud of nationality
No problem Neighbours: People different race	Confidence: Police	Not Justifiable: someone accepting a bribe
No Confidence: Parliament	No problem Neighbours: Homosexuals	Confidence: Police
No problem Neighbours: Homosexuals	No problem Neighbours: Immigrants/foreign workers	Not Justifiable: cheating on taxes
No problem Neighbours: Immigrants/foreign workers	Proud of nationality	Not Justifiable: claiming government benefits
Not Justifiable: claiming government benefits	Not Justifiable: someone accepting a bribe	Not Justifiable: avoiding a fare on public transport
No problem Neighbours: People AIDS	No problem Neighbours: People AIDS	Not Justifiable: suicide
Proud of nationality	Not Justifiable: claiming government benefits	No problem Neighbours: People different race
No Confidence: Civil Services	Not Justifiable: cheating on taxes	Confidence: Armed Forces
Not Justifiable: cheating on taxes	Confidence: Justice System/Courts	Confidence: The Civil Services
Type 2	**Type 2**	**Type 2**
Proud of nationality	No problem Neighbours: People different race	No problem Neighbours: People different race
Not Justifiable: someone accepting a bribe	No Confidence: Parliament	No problem Neighbours: Homosexuals
Confidence: Police	Not Justifiable: someone accepting a bribe	Confidence: Police
No problem Neighbours: People different race	No problem Neighbours: Homosexuals	No problem Neighbours: Immigrants/foreign workers
Not Justifiable: cheating on taxes	No problem Neighbours: People AIDS	Not Justifiable: someone accepting a bribe
Not Justifiable: claiming government benefits	No problem Neighbours: Immigrants/foreign workers	No problem Neighbours: People AIDS
No problem Neighbours: Immigrants/foreign workers	No Confidence: Civil Services	Proud of nationality
Not Justifiable: avoiding a fare on public transport	No Confidence: Major Companies	Not Justifiable: claiming government benefits
Confidence: Armed Forces	Not Justifiable: claiming government benefits	Not Justifiable: cheating on taxes
No problem Neighbours: Homosexuals	No Confidence: The Press	Justifiable: divorce
	Type 3	**Type 3**
	Not Justifiable: someone accepting a bribe	No problem Neighbours: Homosexuals
	Proud of nationality	No problem Neighbours: People different race
	Not Justifiable: cheating on taxes	No Confidence: Parliament
	Not Justifiable: suicide	No problem Neighbours: People AIDS
	Not Justifiable: claiming government benefits	No problem Neighbours: Immigrants/foreign workers
	Not Justifiable: avoiding a fare on public transport	Not Justifiable: someone accepting a bribe
	Not Justifiable: prostitution	No Confidence: Civil Services
	No problem Neighbours: People different race	No Confidence: Churches
	Confidence: Police	No Confidence: Major Companies
	Not Justifiable: abortion	Not Justifiable: claiming government benefits
		Type 4
		Not Justifiable: someone accepting a bribe
		Proud of nationality
		Not Justifiable: claiming government benefits
		Not Justifiable: cheating on taxes
		Not Justifiable: avoiding a fare on public transport
		No Confidence: Parliament
		Not Justifiable: suicide
		No problem Neighbours: People different race
		Not Justifiable: prostitution
		No Confidence: The Press

Notes: This table reports the 10 most important features for a n-type k-means model, where $n \in \{2, 3, 4\}$.

Bibliography

Bibliography

Acemoglu, D., Egorov, G., and Sonin, K. (2013). A political theory of populism. *The Quarterly Journal of Economics*, 128(2):771–805.

Achen, C. H. and Bartels, L. M. (2002). Ignorance and Bliss in Democratic Politics: Party Competition with Uninformed Voters. *Prepared for presentation at the Annual Meeting of the Midwest Political Science Association, Chicago*.

Adena, M., Enikolopov, R., Petrova, M., Santarosa, V., and Zhuravskaya, E. (2015). Radio and the Rise of The Nazis in Prewar Germany. *The Quarterly Journal of Economics*, 130(4):1885–1939.

Airoldi, E. M., Blei, D., Erosheva, E. A., and Fienberg, S. E. (2014). *Handbook of mixed membership models and their applications*. CRC Press.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Alatas, V., Chandrasekhar, A. G., Mobius, M., Olken, B. A., and Paladines, C. When Celebrities Speak: A Nationwide Twitter Experiment Promoting Vaccination In Indonesia: Working Paper.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., and Wacziarg, R. (2003). Fractionalization. *Journal of Economic growth*, 8(2):155–194.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the origins of gender roles: Women and the plough. *Quarterly Journal of Economics*, 128(2):469–530.

Alesina, A., Glaeser, E. L., and Sacerdote, B. (2001). Why Doesn't the United States Have a European-Style Welfare State? *Brookings Papers on Economic Activity*, 2001(2):187–277.

Alesina, A. and La Ferrara, E. (2005). Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, 43:721–761.

Aletras, N., Baldwin, T., Lau, J. H., and Stevenson, M. (2014). Representing topics labels for exploring digital libraries. *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries,IEEE Press*, pages 239–248.

Aletras, N. and Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Algan, Y., Guriev, S., Papaioannou, E., and Passari, E. (2017). The European Trust Crisis and the Rise of Populism. *CEPR Discussion Papers*.

Ali, S. N. and Bénabou, R. Image Versus Information: Changing Societal Norms and Optimal Privacy: Working Paper.

Ali, S. N. and Lin, C. (2013). Why People Vote: Ethical Motives and Social Incentives. *American Economic Journal: Microeconomics*, 5(2):73–98.

Anderson, T. W., Rubin, H., et al. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, 20(1):46–63.

Andrews, I. (2018). Valid Two-Step Identification-Robust Confidence Sets for GMM. *Review of Economics and Statistics*, 100(2):337–348.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak Instruments in IV Regression: Theory and Practice. *Annual Review of Economics*.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ansolabehere, S., Rodden, J., and Snyder, J. M. (2006). Purple america. *Journal of Economic Perspectives*, 20(2):97–118.

Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58(2):277–297.

Arrow, K. J. (2000). Increasing Returns: Historiographic Issues and Path Dependence. *The European Journal of the History of Economic Thought*, 7(2):171–180.

Arthur, W. B. (1989). Competing Technologies, Increasing Returns, and Lock-in by Historical Events. *The Economic Journal*, 99(394):116–131.

Arthur, W. B. (1994). *Increasing Returns and Path Dependence in the Economy*. University of Michigan Press.

Ash, E. The political economy of tax laws in the US states.

Ashok, V., Kuziemko, I., and Washington, E. (2015). Support for redistribution in an age of rising inequality: New stylized facts and some tentative explanations. *National Bureau of Economic Research*.

Atkin, D., Colson-Sihra, E., and Shayo, M. How do we choose our identity? a revealed preference approach using food consumption.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., and Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, 32(3):259–280.

BAMF (2016). Aktuelle Zahlen zu Asyl. *Bundesamt für Migration und Flüchtlinge*.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). Ceo behavior and firm performance. *Journal of Political Economy*, 128(4):1325–1369.

Barberá, P. (2014). How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the US.

Barberá, P. and Rivero, G. (2015). Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review*, 33(6):712–729.

Bartik, T. J. (1991). *Who Benefits from State and Local Economic Development Policies?* Books from Upjohn Press. W.E. Upjohn Institute for Employment Research.

BBC (2017). Social Media Warned to Crack Down on Hate Speech.

BBC (2019). Ilhan Omar: Muslim Lawmaker Sees Rise in Death Threats After Trump Tweet.

Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., and Topalova, P. (2009). Powerful Women: Does Exposure Reduce Bias? *Quarterly Journal of Economics*, 124(4):1497–1540.

Becker, S. O., Fetzer, T., and Novy, D. (2017). Who voted for Brexit? A comprehensive district-level analysis. *Economic Policy*, 32(92):601–650.

Becker, S. O. and Pascali, L. (2019). Religion, Division of Labor, and Conflict: Anti-semitism in Germany over 600 Years. *American Economic Review*, 109(5):1764–1804.

Becker, S. O., Pfaff, S., and Rubin, J. (2016). Causes and Consequences of the Protestant Reformation. *Explorations in Economic History*, 62:1–25.

Bénabou, R. (2008). Ideology. *Journal of the European Economic Association*, 6(2-3):321–352.

Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. *The Review of Economic Studies*, 80(2 (283)):429–462.

Bénabou, R. and Tirole, J. Laws and Norms.

Bénabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, 96(5):1652–1678.

Bertrand, M. and Kamenica, E. Coming apart? Cultural distances in the United States over time.

Bessi, A., Zollo, F., Del Vicario, M., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2015). Trend of Narratives in the Age of Misinformation. *PLOS ONE*, 10(8):1–16.

Bhuller, M., Havnes, T., Leuven, E., and Mogstad, M. (2013). Broadband Internet: An Information Superhighway to Sex Crime? *Review of Economic Studies*, 80(4):1237–1266.

Blanchflower, D. G. and Oswald, A. J. (2008). Is well-being U-shaped over the life cycle? *Social science & medicine (1982)*, 66(8):1733–1749.

Blaydes, L. and Grimmer, J. (2013). Political Cultures: Exploring the Long-Run Determinants of Values Transmission. In *Annual Meeting of the International Political Economy Society, Claremont Graduate School, Claremont, CA*.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *Quarterly Journal of Economics*, 131(4):1753–1794.

Bossert, W., D'Ambrosio, C., and La Ferrara, E. (2011). A generalized index of fractionalization. *Economica*, 78(312):723–750.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2017). Greater Internet Use Is Not Associated With Faster Growth in Political Polarization Among US Demographic Groups. *Proceedings of the National Academy of Sciences*, 114(40):10612–10617.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2020). Cross-Country Trends in Affective Polarization. *National Bureau of Economic Research*.

Brown, B. (2018). The Trump Twitter Archive.

Buisseret, P. and van Weelden, R. (2017). Crashing the Party? Elites, Outsiders, and Elections. In *Wallis conference*.

Bursztyn, L., Cantoni, D., Funk, P., and Yuchtman, N. Polls, the Press, and Political Participation: The Effects of Anticipated Election Closeness on Voter Turnout: NBER Working Papers.

Bursztyn, L., Egorov, G., Enikolopov, R., and Petrova, M. Social Media and Xenophobia: Evidence from Russia: Working Paper.

Bursztyn, L., Egorov, G., and Fiorin, S. From Extreme to Mainstream: How Social Norms Unravel: Working Paper.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. Misperceived Social Norms: Female Labor Force Participation in Saudi Arabia: NBER Working Papers.

Bursztyn, L. and Jensen, R. (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics*, 130(3):1329–1367.

Cagé, J., Hervé, N., and Viaud, M.-L. (2015). The production of information in an online world. *The Review of Economic Studies*.

Canen, N., Kendall, C., and Trebbi, F. (2020). Unbundling polarization. *Econometrica*, 88(3):1197–1233.

Card, D. and Dahl, G. B. (2011). Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior. *The Quarterly Journal of Economics*, 126(1):103–143.

Caselli, F. and Coleman, W. J. (2013). On the theory of ethnic conflict. *Journal of the European Economic Association*, 11(suppl. 1):161–192.

Center, P. R. (2018). News Use Across Social Media Platforms 2018. *Available at: http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/.*

Chan, J., Ghose, A., and Seamans, R. (2016). The Internet and Racial Hate Crime: Offline Spillovers from Online Access. *MIS Quarterly*, 40(2):381–403.

Chang, J., Boyd-Graber, J., Blei, D. M., Wang, C., and Gerrish, S. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, pages 288–296.

Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., and Camerer, C. (2018). Econographics. *National Bureau of Economic Research.*

Che, Y., Lu, Y., Pierce, J. R., Schott, P. K., and Tao, Z. Does trade liberalization with China influence US elections?

Chetty, R., Hendren, N., Jones, M. R., and Porter, S. R. (2018). Race and economic opportunity in the united states: An intergenerational perspective. Working Paper 24441, National Bureau of Economic Research.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., and Winter, P. L. (2006). Managing Social Norms for Persuasive Impact. *Social Influence*, 1(1):3–15.

Colella, F., Lalive, R., Sakalli, S. O., and Thoenig, M. (2019). Inference with Arbitrary Clustering. IZA Discussion Papers 12584, Institute of Labor Economics (IZA).

Colussi, T., Isphording, I. E., and Pestel, N. (2016). Minority Salience and Political Extremism.

Dahl, G. and DellaVigna, S. (2009). Does Movie Violence Increase Violent Crime? *The Quarterly Journal of Economics*, 124(2):677–734.

Dal Bó, E., Dal Bó, P., and Eyster, E. (2018). The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects. *The Review of Economic Studies*, 85(2):964–998.

Dal Bó, E., Finan, F., Folke, O., Rickne, J., and Persson, T. Economic losers and political winners: The rise of the radical right in Sweden.

Daughety, A. F. and Reinganum, J. F. (2010). Public Goods, Social Pressure, and the Choice between Privacy and Publicity. *American Economic Journal: Microeconomics*, 2(2):191–221.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The Spreading of Misinformation Online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

DellaVigna, S., Enikolopov, R., Mironova, V., Petrova, M., and Zhuravskaya, E. (2014). Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia. *American Economic Journal: Applied Economics*, 6(3):103–132.

DellaVigna, S. and Gentzkow, M. (2010). Persuasion: Empirical Evidence. *Annual Review of Economics*, 2(1):643–669.

DellaVigna, S. and La Ferrara, E. Economic and Social Impacts of the Media: NBER Working Papers.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2016). Voting to Tell Others. *The Review of Economic Studies*, 84(1):143–181.

Desmet, K. and Wacziarg, R. (2018). The cultural divide. *National Bureau of Economic Research*.

Dinar, C., Mair, T., Rafael, S., Rathje, J., and Schramm, J. (2016). Hetze gegen Flüchtlinge in Sozialen Medien. *Amadeu Antonio Stiftung*.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29.

Dippel, C., Gold, R., and Heblich, S. Globalization and its (dis-) content: Trade shocks and voting behavior.

Dorn, D., Hanson, G., Majlesi, K., et al. Importing political polarization? The electoral consequences of rising trade exposure.

Downs, A. (1957). An economic theory of political action in a democracy. *Journal of Political Economy*, 65(2):135–150.

Draca, M. and Schwarz, C. (2018). How Polarized Are Citizens? Measuring Ideology from the Ground-up.

Duclos, J.-Y., Esteban, J., and Ray, D. (2004). Polarization: Concepts, measurement, estimation. *Econometrica*, 72(6):1737–1772.

Durante, R. and Zhuravskaya, E. (2018). Attack When the World Is Not Watching? US News and the Israeli-Palestinian Conflict. *Journal of Political Economy*, 126(3):1085–1133.

Economist Intelligence Unit (2020). Democracy Index 2019.

Edwards, B. T. (2018). Trump from Reality TV to Twitter, or the Selfie-Determination of Nations. *Arizona Quarterly: A Journal of American Literature, Culture, and Theory*, 74(3):25–45.

Eisensee, T. and Strömberg, D. (2007). News Droughts, News Floods, and U. S. Disaster Relief. *The Quarterly Journal of Economics*, 122(2):693–728.

Enikolopov, R., Makarin, A., and Petrova, M. (2020). Social media and protest participation: Evidence from russia. *Econometrica*, 88(4):1479–1514.

Enke, B. (2020). Moral values and voting. *Journal of Political Economy*, 128(10).

Esteban, J.-M. and Ray, D. (1994). On the measurement of polarization. *Econometrica*, pages 819–851.

FBI (2015). Hate Crime Data Collection Guidelines And Training Manual. *Criminal Justice Information Services (CJIS) Division Uniform Crime Reporting (UCR) Program.*

Fiorina, M. P. and Abrams, S. J. (2008). Political polarization in the American public. *Annual Review of Political Science*, 11:563–588.

Fouka, V. and Voth, H.-J. Reprisals Remembered: German-Greek Conflict and Car Sales during the Euro Crisis: CEPR Discussion Papers.

Fukuyama, F. (1992). *The end of history and the last man*. Simon and Schuster.

Gabler, N. (2016). The Internet and Social Media Are Increasingly Divisive and Undermining of Democracy. *Alternet*.

Gavazza, A., Nardotto, M., and Valletti, T. (2018). Internet and Politics: Evidence from U.K. Local Elections and Local Government Policies. *The Review of Economic Studies*, 86(5):2092–2135.

Gawker (2007). Twitter Blows Up at SXSW Conference.

Gennaioli, N. and Tabellini, G. (2019). Identity, Beliefs, and Political Conflict. *Available at SSRN 3300726*.

Gentzkow, M. (2006). Television and Voter Turnout. *The Quarterly Journal of Economics*, 121(3):931–972.

Gentzkow, M. (2016). Polarization in 2016. *Toulouse Network of Information Technology white paper*.

Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340.

Geoghegan, V. (2003). *Political ideologies: An introduction*. Routledge.

Gerber, A. S., Green, D. P., and Larimer, C. W. (2008). Social Pressure and Voter Turnout: Evidence from a Large-scale Field Experiment. *American Political science review*, 102(1):33–48.

Glaeser, E. L. and Ward, B. A. (2006). Myths and realities of American political geography. *Journal of Economic Perspectives*, 20(2):119–144.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624.

Gould, E. D. and Klor, E. F. (2016). The Long-run Effect of 9/11: Terrorism, Backlash, and the Assimilation of Muslim Immigrants in the West. *The Economic Journal*, 126(597):2064–2114.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Grimmer, J. (2009). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35.

Gross, J. H. and Manrique-Vallier, D. (2012). A mixed-membership approach to the assessment of political ideology from survey responses. In *Individual Presentation, Society for Political Methodology, 29th Annual Summer Meeting, Chapel Hill, NC*.

Grossman, G. M. and Helpman, E. Identity politics and trade policy.

Guardian, T. (2017). CPS to Crack Down on Social Media Hate Crime, says Alison Saunders, by Vikram Dodd.

Guess, A., Nyhan, B., and Reifler, J. (2018). Selective Exposure to Misinformation: Evidence from the Consumption of Fake News during the 2016 U.S. Presidential Campaign. *Working Paper*.

Guess, A. M. (2018). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *Working Paper*.

Guiso, L., Herrera, H., Morelli, M., and Sonno, T. (2017). Demand and Supply of Populism. *IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University*.

Guriev, S., Melnikov, N., and Zhuravskaya, E. (2019). 3g internet and confidence in government. *Available at SSRN 3456747*.

Hanes, E. and Machin, S. (2014). Hate Crime in the Wake of Terror Attacks: Evidence from 7/7 and 9/11. *Journal of Contemporary Criminal Justice*, 30(3):247–267.

Hansen, S., McMahon, M., and Prat, A. (2014). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics*.

Haustein, S. and Costas, R. (2014). Determining Twitter Audiences: Geolocation and Number of Followers. *ALM*, 4:6.

Healy, A. and Malhotra, N. (2013). Retrospective voting reconsidered. *Annual Review of Political Science*, 16:285–306.

Hobbs, W. and Lajevardi, N. (2019). Effects of Divisive Political Campaigns on the Day-to-Day Segregation of Arab and Muslim Americans. *American Political science review*, 113(1):270–276.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *the Journal of machine Learning research*, 14(1):1303–1347.

Hölig, S. and Hasebrink, U. (2016). *Reuters Institute Digital News Survey 2017: Ergebnisse für Deutschland*, volume Nr. 38 of *Arbeitspapiere des Hans-Bredow-Instituts*. Verlag Hans-Bredow-Institut, Hamburg.

Hopkins, D. J. and Washington, S. (2019). The Rise of Trump, the Fall of Prejudice? Tracking White Americans' Racial Attitudes 2008-2018 via a Panel Survey. *Working Paper*.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton University Press.

Inglehart, R. F., Basanez, M., and Moreno, A. (2010). *Human Values and Beliefs: A Cross-Cultural Sourcebook*. University of Michigan Press, Ann Arbor.

Jelveh, Z., Kogut, B., and Naidu, S. (2015). Political language in economics. *Working Paper*.

Jensen, J., Naidu, S., Kaplan, E., Wilse-Samson, L., Gergen, D., Zuckerman, M., and Spirling, A. (2012). Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech [with Comments and Discussion]. *Brookings Papers on Economic Activity*, Fall:1–81.

Jha, S. (2013). Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia. *American Political science review*, 107(4):806–832.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kamps, H. J. (2015). Who Are Twitter's Verified Users?

Kaplan, E., Spenkuch, J. L., and Sullivan, R. (2019). Measuring Geographic Polarization: Theory and Long-Run Evidence. *Working Paper*.

Kim, D. and Oh, A. (2011). Topic chains for understanding a news corpus. *International Conference on Intelligent Text Processing and Computational Linguistics, Springer*, pages 163–176.

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., and Morstatter, F. (2017). Archiving Information from Geotagged Tweets to Promote Reproducibility and Comparability in Social Media Research. *Big Data & Society*, 4(2):2053951717736336.

Kreißel, P., Ebner, J., Urban, A., and Guhl, J. (2018). Hass auf Knopfdruck: Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz. *Institute for Strategic Dialogue*.

Kuran, T. (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.

Lau, J. H. and Baldwin, T. (2016). The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. *Proceedings of NAACL-HLT*, pages 483–487.

Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics*, pages 1536–1545.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings ofthe 14th Conference ofthe European Chapter ofthe Association for Computational Linguistic*, pages 530–539.

Leigh, A. (2009). Does the world economy swing national elections? *Oxford Bulletin of Economics and Statistics*, 71(2):163–181.

Liebowitz, S. J. and Margolis, S. E. (1999). Path Dependence. *Encyclopedia of law and economics*.

Manacorda, M. and Tesei, A. Liberation Technology: Mobile Phones and Political Mobilization in Africa: CEPR Discussion Papers.

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies*, 60(3):531–542.

Martin, G. J. and Yurukoglu, A. (2017). Bias in Cable News: Persuasion and Polarization. *American economic review*, 107(9):2565–2599.

Matz, S. C., Kosinski, M., Nave, G., and Stillwell, D. J. (2017). Psychological Targeting as an Effective Approach to Digital Mass Persuasion. *Proceedings of the National Academy of Sciences*, 114(48):12714–12719.

Miller, C. and Smith, J. (2017). Anti-Islamic Content on Twitter. *Centre for the Analysis of Social Media at Demos*.

Miller, D. T. and Prentice, D. A. (1994). Collective Errors and Errors about the Collective. *Personality and Social Psychology Bulletin*, 20(5):541–550.

Montalvo, J. G. and Reynal-Querol, M. (2005). Ethnic polarization, potential conflict, and civil wars. *American economic review*, 95(3):796–816.

Mudde, C. (2004). The populist zeitgeist. *Government and opposition*, 39(4):541–563.

Mukand, S. and Rodrik, D. The Political Economy of Ideas: On Ideas Versus Interests in Policymaking: Working Paper.

Müller, K. and Schwarz, C. (2018a). Fanning the Flames of Hate: Social Media and Hate Crime. *Working Paper*.

Müller, K. and Schwarz, C. (2018b). From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment. *Available at SSRN: https://ssrn.com/abstract=3149103*.

Munro, E. and Ng, S. (2019). Latent Dirichlet Analysis of Categorical Survey Responses. *arXiv preprint arXiv:1910.04883*.

Murray, D. (2016). Populism: It's the BBC's new buzzword, being used to sneer at the 'uneducated' 17 million who voted for Brexit. *Daily Mail*, 16:December.

New York Times (2018). The Man Behind the President's Tweets.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 100–108.

News, N. B. (2017). Advocates Warn of Possible Underreporting in FBI Hate Crime Data, by Chris Fuchs.

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*, 49(6):1417–1426.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574.

Norris, P. (2016). *Democratic deficit: Critical citizens revisited*. Cambridge University Press, Cambridge [etc.].

Norris, P. and Inglehart, R. (2019). *Cultural backlash: Trump, Brexit, and authoritarian populism*. Cambridge University Press.

NYT (2019). Tracking Trump's Visits to His Branded Properties.

OECD (2016). Broadband Statistics.

Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M., and Räsänen, P. (2014). Exposure to Online Hate Among Young Social Media Users. *Soul of society: a focus on the lives of children & youth*, 18:253–273.

Olea, J. L. M. and Pflueger, C. (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics*, 31(3):358–369.

Ortoleva, P. and Snowberg, E. (2015). Overconfidence in political behavior. *American economic review*, 105(2):504–535.

Ott, C. and Gür-Seker, D. (2019). Rechtspopulismus und Social Media: Wie Wortgebräuche in Social Media sprachkritisch betrachtet werden können. In *Soziale Medien in Schule und Hochschule: Linguistische, sprach- und mediendidaktische Perspektiven*, pages 279–318. Peter Lang AG.

Panagopoulos, C. (2006). The Polls-Trends: Arab and Muslim Americans and Islam in the aftermath of 9/11. *International Journal of Public Opinion Quarterly*, 70(4):608–624.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding From You*. Penguin UK.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Perez-Truglia, R. and Cruces, G. (2017). Partisan Interactions: Evidence from a Field Experiment in the United States. *Journal of Political Economy*, 125(4):1208–1243.

Pesaran, M. H. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica*, 74(4):967–1012.

Petrova, M., Sen, A., and Yildirim, P. (2017). Social Media and Political Donations: New Technology and Incumbency Advantage in the United States. *Working Paper*.

Piketty, T. (2018). Brahmin Left versus Merchant Right: Rising Inequality and the Changing Structure of Political Conflict. *WID Working Paper Series*.

Poole, K. T. and Rosenthal, H. (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.

ProPublica (2017). Why America Fails at Gathering Hate Crime Statistics, by Ken Schwencke.

Putthividhy, D., Attias, H. T., and Nagarajan, S. S. (2010). Topic regression multi-modal latent dirichlet allocation for image annotation. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3408–3415.

Rodrik, D. Populism and the Economics of Globalization.

Rooduijn, M., van Kessel, S., Froio, C., Pirro, A., de Lange, S., Halikiopoulou, D., Lewis, P., Mudde, C., and Taggart, P. (2019). The PopuList: An Overview of Populist, Far Right, Far Left and Eurosceptic Parties in Europe. *The PopuList. https://popu-list. org*.

Schmidt, A. L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2017). Anatomy of News Consumption on Facebook. *Proceedings of the National Academy of Sciences*, 114(12):3035–3039.

Schwarz, C. (2018). ldagibbs: A command for topic modeling in Stata using latent Dirichlet allocation. *Stata Journal*, 18(1):101–117.

Shayo, M. (2009). A model of social identity with an application to political economy: Nation, class, and redistribution. *American Political science review*, 103(2):147–174.

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292.

Stephens-Davidowitz, S. (2014). The Cost of Racial Animus on a Black Candidate: Evidence using Google Search Data. *Journal of Public Economics*, 118(C):26–40.

Stock, J. and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In Donald W.K. Andrews , editor, *Identification and Inference for Econometric Models*, pages 80–108. Cambridge University Press, New York.

Sun, L. (2018). Implementing Valid Two-Step Identification-Robust Confidence Sets For Linear Instrumental-Variables Models. *The Stata Journal*, 18(4):803–825.

Sunstein, C. R. (2009). *Republic.com 2.0*. Princeton University Press.

Sunstein, C. R. (2017). *# Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Takhteyev, Y., Gruzd, A., and Wellman, B. (2012). Geography of Twitter Networks. *Social networks*, 34(1):73–81.

Times, F. (2017a). Powerhouse Germany Badly Trailing Rivals in Broadband.

Times, N. Y. (2016). How Facebook Warps Our Worlds, By Frank Bruni.

Times, N. Y. (2017b). How Fiction Becomes Fact on Social Media, By Benedict Carey.

Times, N. Y. (2017c). Seeking Asylum in Germany, and Finding Hatred, By Ainara Tiefenthäler, Shane O'neill and Andrew Michael Ellis.

Times, N. Y. (2017d). Trump Shares Inflammatory Anti-Muslim Videos, and Britain's Leader Condemns Them, By Peter Baker and Eeileen Sullivan.

Tupes, E. C. and Christal, R. E. (1961). Stability of personality factors based on trait ratings. *USAF ASD Tech. Rep.*

Vincent, A. (2009). *Modern political ideologies*. John Wiley & Sons.

Voigtlander, N. and Voth, H.-J. (2012). Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany. *The Quarterly Journal of Economics*, 127(3):1339–1392.

Voigtlander, N. and Voth, H.-J. (2015). Nazi Indoctrination and Anti-Semitic Beliefs in Germany. *Proceedings of the National Academy of Sciences of the United States of America*, 112(26):7931–7936.

Wang, C., Blei, D. M., and Heckerman, D. (2008). Continuous Time Dynamic Topic Models. *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 579–586.

Wang, Y. S., Matsueda, R. L., Erosheva, E. A., et al. (2017). A variational EM method for mixed membership models with multivariate rank data: An analysis of public policy preferences. *The Annals of Applied Statistics*, 11(3):1452–1480.

Weber, M. (1919). *Politik als Beruf*. Vier Vorträge vor dem Freistudentischen Bund. Zweiter Vortrag.

Westfall, J., van Boven, L., Chambers, J. R., and Judd, C. M. (2015). Perceiving Political Polarization in the United States: Party Identity Strength and Attitude Extremity Exacerbate the Perceived Partisan Divide. *Perspectives on Psychological Science*, 10(2):145–158.

Willnat, L., Weaver, D. H., and Wilhoit, G. C. (2019). The American Journalist in the Digital Age. *Journalism Studies*, 20(3):423–441.

Wolfers, J. (2002). Are voters rational? Evidence from gubernatorial elections. *Work. pap., The Wharton School*.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data*, volume 1 of *MIT Press Books*. The MIT Press.

Yanagizawa-Drott, D. (2014). Propaganda and Conflict: Evidence from the Rwandan Genocide. *The Quarterly Journal of Economics*, 129(4):1947–1994.